David Fernandes Semedo

Master of Computer Science and Engineering

Bridging Vision and Language over Time with Neural Cross-modal Embeddings

Thesis submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy in
Computer Science

Adviser: João Miguel da Costa Magalhães, Associate Professor, Faculdade de Ciências e Tecnologia, Universidade NOVA de Lisboa

Examination Committee

Chair:	Professor Nuno Manuel Robalo Correia
Rapporteurs:	Professor Cees Snoek
	Professor Richang Hong
Members:	Professor Nuno Manuel Robalo Correia
	Professor João Miguel da Costa Magalhães
	Professor Benoit Huet
	Professor Ludwig Krippahl
	Dr. Ricardo Gamelas Sousa

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

March, 2020

This document was created using the (pdf)LaTeX processor, based on the "novathesis" template[1], developed at the Dep. Informática of FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

In memory of my mother (1966-2017).
"Let the wind carry you home".

Acknowledgements

First, I would like to express my deepest gratitude to my advisor, João Magalhães. This was a truly enriching academic journey, where his guidance and utmost predisposition to discuss and pursue new ideas were key to this thesis. I am very thankful for all the support, encouragement and wise mentorship. What started as an academic relationship, I am sure now became a friendship. I also would like to thank my thesis advisory committee members, Benoit Huet and Ludwig Krippahl, for the insightful and constructive feedback provided, both on low-level and high-level aspects of the work, that contributed a lot to this thesis. Moreover, I thank Cees Snoek, Richang Hong and Ricardo Sousa, for the great and insightful comments and questions, raised on the thesis defense.

Second, I want to thank my host institution, Departamento de Informática from Universidade NOVA de Lisboa, for supporting my research and providing me all the resources needed. A special thanks to the Ph.D. program coordinator, Nuno Correia, for the precious help with all the administrative stuff. I also would like to thank the sources that funded and made this thesis possible, the CMU Portugal research project GoLocal Ref. CMUP-ERI/TIC/0046/2014, the H2020 ICT project COGNITUS with the grant agreement n^{o} 687605 and the FCT project NOVA LINCS Ref. UID/CEC/04516/2019. I also gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPUs used for this research.

To all the folks from P3-13 Lab, whom I shared great moments throughout these 4 years: Flávio Martins, André Mourão, Pedro Albuquerque, Rui Rodrigues, Camila Wohlmuth, Rui Madeira, Gustavo Gonçalves, Carla Viegas and Rui Nóbrega. Thank you all for the warm environment in the lab, in which we shared both stressful and relaxed moments. A special thanks to Flávio, for our great *hackathons* and for always being available to help me with cluster-related stuff.

To my longtime friends, Carlos Crespo and João Mirão, for their friendship, for the moments we spent together, and for all the beers. To André and Flávia for all the support and joyful moments. To Olga, Pedro, Ana, José, princess Inês, Lurdes, Beatriz and Nuno, for all the love and support. All of you supported me in ways that you cannot imagine.

To all my family, specially to my grandparents, who were always there to help me. To my parents for the education, for believing in me and giving me infinite support.

Specifically, to my father, for being my inspiration, and to my mother, for all her guidance. I know that her spirit was and will be eternally present in my life. To my sister Catarina, which whom I shared marvelous moments, doing our silly things.

The deepest and warm thanks to my beloved life partner, Marta Rolho. Words are not enough to express her importance. Thank you for always being there, for calming me down in the most stressful moments, and specially, for being part of my life. Without all the unconditional support, love, encouragement and friendship she provided me, I would not have finished this thesis. You are my future.

Abstract

Giving computers the ability to understand multimedia content is one of the goals of Artificial Intelligence systems. While humans excel at this task, it remains a challenge, requiring bridging vision and language, which inherently have heterogeneous computational representations. Cross-modal embeddings are used to tackle this challenge, by learning a common space that unifies these representations. However, to grasp the semantics of an image, one must look beyond the pixels and consider its **semantic and temporal context**, with the latter being defined by images' textual descriptions and time dimension, respectively. As such, external causes (*e.g.* emerging events) change the way humans interpret and describe the same visual element over time, leading to the evolution of visual-textual correlations.

In this thesis we investigate models that capture patterns of visual and textual interactions over time, by incorporating time in cross-modal embeddings: **1)** in a **relative manner**, where by using pairwise temporal correlations to aid data structuring, we obtained a model that provides better visual-textual correspondences on dynamic corpora, and **2)** in a **diachronic manner**, where the temporal dimension is fully preserved, thus capturing visual-textual correlations evolution under a principled approach that jointly models vision+language+time. Rich insights stemming from data evolution were extracted from a 20 years large-scale dataset. Additionally, towards improving the effectiveness of these embedding learning models, we proposed a novel loss function that increases the expressiveness of the standard triplet-loss, by making it adaptive to the data at hand. With our **adaptive triplet-loss**, in which triplet specific constraints are inferred and scheduled, we achieved state-of-the-art performance on the standard cross-modal retrieval task.

Keywords: Temporal Embeddings; Cross-modal embeddings; Multimedia understanding; Vision and language; Neural networks

Resumo

Providenciar aos computadores a habilidade de compreender conteúdo multimédia, é um dos objectivos de sistemas de Inteligência Artificial. Enquanto que os humanos apresentam um desempenho notável nesta tarefa, ainda constitui um desafio que quer requer a unificação de visão e linguagem, que inerentemente têm representações computacionais heterogéneas. Embeddings Cross-modais são utilizados para atacar este desafio, aprendendo um espaço comum que unifica estas representações. No entanto, para compreender a semântica de uma imagem, é necessário olhar além dos seus pixeis e considerar o seu contexto semântico e temporal, sendo este definido pelas legendas e dimensão temporal, respectivamente. Como tal, causas externas (e.g. eventos emergentes) alteram a forma como os humanos interpretam e descrevem o mesmo elemento visual ao longo do tempo, originando evolução das correlações visuais-textuais.

Nesta tese investigamos métodos que capturam padrões de interacção entre visão e linguagem ao longo do tempo, através da incorporação do tempo em embeddings cross-modais: **1)** de **forma relativa**, em que utilizando correlações temporais de pares para estruturar os dados, obtivemos um modelo que melhor captura correspondências visuais-textuais em colecções dinâmicas, e **2)** de uma **forma diacrónica**, em que a dimensão temporal é preservada, capturando a evolução de correlações visuais-textuais, seguindo uma abordagem bem fundamentada que modela conjuntamente visão+linguagem+tempo. Intuições bastante ricas, com origem na evolução dos dados, foram extraídas de um dataset de larga escala (20 anos). Adicionalmente, para melhorar a eficácia de modelos de aprendizagem de embeddings, propusemos uma nova função de custo que aumenta a expressividade da triplet-loss standard, tornando-a adaptativa. Com a **triplet-loss adaptativa**, em que restrições específicas a cada triplo são inferidas e escalonadas, obtivemos uma performance ao nível do estado da arte na tarefa de pesquisa cross-modal.

Palavras-chave: Embeddings Temporais; Embeddings Cross-modal; Compreensão de Multimedia; Visão e Linguagem; Redes Neuronais

Contents

List of Figures xvii

List of Tables xxi

Acronyms xxiii

Symbols xxv

1 Introduction 1

1.1 Bridging Vision and Language over Time . . . 1

1.1.1 Cross-modal Embeddings to Bridge Vision and Language . . . 3

1.1.2 Going Beyond the Static Corpus Assumption . . . 4

1.1.3 Problem Framework Formal Definition . . . 5

1.1.4 Improving Neural Cross-modal Embedding Learning . . . 7

1.2 Objectives and Research Questions . . . 8

1.2.1 Scheduled learning of Adaptive Triplet Loss . . . 8

1.2.2 Temporal Cross-modal Structuring . . . 9

1.2.3 Diachronic Cross-modal Structuring . . . 10

1.3 Contributions and Impact . . . 10

1.3.1 Research Publications . . . 11

1.3.2 Multimodal Research Benchmark Datasets . . . 12

1.3.3 Industry Impact and Use-Cases . . . 13

1.4 Document Organization . . . 14

2 Related Work 15

2.1 Feature Representations for Images and Text . . . 15

2.1.1 Image Representations . . . 15

2.1.2 Text Representations . . . 19

2.2 Computationally Bridging Vision and Language . . . 21

2.2.1 Multimodal Modeling of Images and Text . . . 23

2.2.2 Cross-modal Embedding Learning . . . 24

2.2.3 Cross-modal Projection Functions . . . 26
2.2.4 Neural Cross-modal Embeddings Learning . . . 29
2.2.5 Cross-modal Loss Functions and Optimization . . . 34
2.3 Modeling Temporal Evidence . . . 40
2.3.1 Capturing and Representing Temporal Clues . . . 40
2.3.2 Modeling Data Evolution . . . 42
2.4 Evaluation Metrics . . . 45

3 Scheduled Adaptive Margin for Neural Cross-Modal Embeddings 49
3.1 Cross-modal Embedding Space Structure Definition . . . 51
3.1.1 Embedding Properties . . . 52
3.2 Adaptive Embedding Learning . . . 52
3.2.1 Static Maximum-margin Formulation . . . 53
3.2.2 Limitations of Standard Triplet Loss on Neural Models . . . 53
3.2.3 Adaptive Triplet Loss Formulation . . . 55
3.3 Scheduled Activation of Adaptive Margins . . . 56
3.3.1 Scheduler Function . . . 57
3.3.2 Adaptive Margin . . . 57
3.3.3 Neural Model and Architecture . . . 59
3.4 Optimization and Triplet Sampling . . . 60
3.5 Evaluation . . . 61
3.5.1 Datasets . . . 61
3.5.2 Methodology . . . 62
3.5.3 Training and Implementation Details . . . 63
3.6 Results and Discussion . . . 63
3.6.1 Cross-modal Retrieval . . . 63
3.6.2 Scheduled Adaptive Margins Analysis . . . 66
3.6.3 Analysis of Activation Phase f_a and λ Impact . . . 69
3.6.4 Qualitative Analysis . . . 70
3.7 Critical Summary . . . 71

4 Temporal Cross-modal Embeddings 77
4.1 Formulating the Temporal Embedding Space Hypothesis . . . 78
4.1.1 Modeling Relative Temporal Correlation . . . 79
4.2 Embedding Definition . . . 81
4.2.1 Temporal Cross-modal Space . . . 81
4.2.2 Time-sensitive Cross-modal Neural Projections . . . 82
4.2.3 Embedding Properties . . . 83

4.3 Temporal Embedding Model Design and Learning . . . 84
4.3.1 Joint Temporal Triplet Ranking Loss . . . 84
4.3.2 Temporal Cross-modal Soft-Constraints . . . 86
4.3.3 Cross-modality similarity . . . 87
4.4 Temporal Soft-Smoothing Correlation Functions . . . 88
4.4.1 Recency-based Correlations . . . 88
4.4.2 Category-based Correlations . . . 88
4.4.3 Topic-based Correlations . . . 89
4.5 Neural Model and Architecture . . . 91
4.6 Evaluation . . . 91
4.6.1 Datasets . . . 92
4.6.2 Methodology . . . 94
4.6.3 Training and Implementation Details . . . 94
4.7 Experiments and Results . . . 95
4.7.1 Cross-Modal Retrieval . . . 95
4.7.2 Media temporal correlations . . . 99
4.8 Critical Summary . . . 102

5 Diachronic Cross-modal Embeddings 103
5.1 Formulating the Diachronic Embedding Space Hypothesis . . . 104
5.2 Embedding Definition . . . 106
5.2.1 Diachronic Cross-modal Space . . . 106
5.2.2 Temporal Embeddings Alignment . . . 107
5.2.3 Time-preserving Projections . . . 108
5.2.4 Embedding Properties . . . 108
5.3 Diachronic Embedding Model Design and Learning . . . 109
5.3.1 From Projections to Triplet Ranking Loss . . . 109
5.3.2 Joint Diachronic Triplet Ranking Loss . . . 110
5.3.3 Binned Structure . . . 110
5.3.4 Continuous Diachronic Structure . . . 113
5.4 Evaluation . . . 117
5.4.1 Dataset - A 20 years Flickr Images Dataset . . . 118
5.4.2 Methodology . . . 118
5.4.3 Training and Implementation Details . . . 120
5.5 Experiments and Results . . . 120
5.5.1 Time Period based Inference . . . 121
5.5.2 Semantic Dispersion over Time . . . 122
5.5.3 Diachronic Semantic Alignment . . . 124

5.5.4 Assessing the Preservation of Temporal Locality Biases 127
5.5.5 Cross-modal Evolution 128
5.6 Critical Summary 131

6 Conclusions and Future Work **133**
6.1 Temporal Information on Cross-modal Embeddings 133
6.2 Learning of Neural Cross-modal Embedding Models 135
6.3 Limitations 136
6.3.1 Adaptive Margins 136
6.3.2 Temporal Cross-modal Embeddings 137
6.3.3 Diachronic Cross-modal embeddings 137
6.4 Future work 138
6.5 Forthcoming Challenges of Multimedia Understanding 140

Bibliography **143**

List of Figures

1.1 Illustration of *diachronicity* in images: Image depicting the same content, *wreckage*, but with different temporal context, are described differently. . . 2
1.2 Illustration of Cross-Modal Embeddings approach. 4
1.3 Relative vs. Absolute Time dimension Modeling. 5
1.4 Triplet loss illustration. Hinge-loss constraints are enforced over triplets to structure the embedding space. 7
1.5 H2020 COGNITUS Project - Providing users a "More like being there"Experience, with User Generated Content. 13

2.1 Overall Architectural Scheme of Convolutional Neural networks. Adapted from [79] 16
2.2 VGG-16 Convolutional Neural Network Architecture. Adapted from [30]. . 17
2.3 Inception module from GoogleNet. Adapted from [115] 17
2.4 Residual learning building block. Adapted from [102]. 18
2.5 Illustration of Bag-of-Word Representation for Sentences. Adapted from [6]. 20
2.6 Example of a multimodal document d^i comprised by an image $\mathbf{x}_V^i$ and a text $\mathbf{x}_T^i$. Illustrates the semantic gap between images and their associated text. . 21
2.7 Contrasting between Cross-modal Learning approaches. Adapted from [91]. 24
2.8 Illustration of Cross-modal Embedding Space. 26
2.9 Multimodal neural architectures proposed in [88]. Both images were adapted from [88]. 31
2.10 Neural architecture of Deep Correlation Canonical Analysis. Adapted from [2]. 32
2.11 Neural architecture of Deep Correlation Canonical Analysis, including the modifications proposed in [133] to deal with overfitting. Adapted from [133]. 32
2.12 Neural architecture of Correspondence Cross-modal Autoencoder. Adapted from [28]. 33
2.13 Neural architecture of Correspondence Full-modal Autoencoder. Adapted from [28]. 33

2.14 Illustration of different a training batch sampling schemes with mini-batch size of $b = 6$. Red edges and blue edges represent positive and negative instances, respectively. 39
2.15 Graphical Representation of D-LDA, for three time slices. Each time slice corresponds to an LDA model. Source [15]. 44
2.16 Visualization of word shifts across time, based on their similarity with other words under a diachronic word embedding. Source [41]. 45

3.1 Adaptive margin constraints are scheduled to be progressively enforced during the training phase. 51
3.2 SAM model architecture. The model is composed by two sub-networks coupled by the loss function $\mathscr{L}_{SAM}$. At each learning epoch t the loss $\mathscr{L}_{SAM}$ imposes triplet-specific constraints, enforcing cluster formation/preservation and organizing instances according to their semantic similarity. 55
3.3 Plot of $\alpha(t)$ with $n_e = 50$. The scheduling training enables a smooth transition from static margins to adaptive margins. 57
3.4 Global average adaptive margin f_m over training epochs (t) on the NUS-WIDE-10k. The left y-axis corresponds to the f_m value and the right y-axis to the scheduling function $\alpha(t)$ value. 67
3.5 t-SNE projections - Scheduled Adaptive Margins between 3 categories. . . 68
3.6 Analysis of the margin values over each epoch (t), between three categories. 68
3.7 Average per-category margin for each category, at each training epoch (t). Average value of f_m between every instance d^i, against all instances d^n of other categories, on NUS-WIDE-10k. 73
3.8 Parameter Analysis (λ and activation function f_a) on Pascal Sentences dataset. 74
3.9 t-SNE Visualization of test instances projections of the NUS-WIDE-10k dataset, on the obtained embedding space. Triangles and circles refer to image and text elements, respectively. Best viewed in color. 74
3.10 Results for query X in the $T \mapsto I$ task. Green border for correct and red for incorrect. 75
3.11 Results for query X in the $I \mapsto T$ task. Green border for correct and red for incorrect. 76

4.1 Temporal dynamics of content from the semantic category *Crash* (Tour-de-France 2016), and temporal pairwise variations with corresponding visual elements. 77
4.2 Temporal dynamics of semantic category *Crash* (TDF2016), and temporal pairwise variations with corresponding visual elements. 79

4.3 Temporal correlations of same-category multimodal data (on the left) follow an unknown density distribution. The temporal cross-modal embedding (on the right) captures these temporal correlations by organizing projected data accordingly, for each specific semantic category. 80
4.4 Temporal cross-modal embedding learning overview. Visual (blue) and textual (purple) instances are mapped to a D dimensional cross-modal space. . 82
4.5 In terms of intra-category structuring, the space is perturbed to approximate temporally correlated instances, and to separate uncorrelated ones. 84
4.6 Constraints violations rationale. 86
4.7 Words temporal relevance. Each plot depicts the mean latent-topical temporal curve ϕ_w, over each day, on the Edinburgh Festival dataset. Vertical lines mark the event timespan. 90
4.8 Cross-modal retrieval *mAP* results, average of $I \mapsto T$ and $T \mapsto I$, per category, on EdFest2016. 99
4.9 Cross-modal retrieval *mAP* results, average of $I \mapsto T$ and $T \mapsto I$, per category, on TDF2016. 99
4.10 Precision-Scope curves for Edinburgh Festival 2016. 100
4.11 Precision-Scope curves for Tour-de-France 2016. 100
4.12 Qualitative analysis of the different temporal correlations on the EdFest2016 and TDF2016 dataset. Each plot depicts the temporal distribution of ground-truth instances, from the categories *Castle* and *Crash*. We use *days* as time granularity. 100
4.13 Temporal vs. Non-Temporal method. 101

5.1 Diachronic Cross-modal Embeddings illustration. 104
5.2 Diachronic cross-modal architecture overview. Visual (blue) and textual (purple) instances, at an instant ts^i, are mapped to a D dimensional diachronic embedding space. A shared temporal structuring layer takes the timestamp ts^i as input and learns an embedding for ts^i, that is then used to independently condition modality projections on time. A diachronic triplet ranking loss is responsible for structuring instances over time. Best viewed in color. 116
5.3 Temporal distribution of the full dataset. The x-axis shows the years while the y-axis shows the number of instances (log-scale). The red dashed vertical line delimits the cut performed due to low number of instances. 117
5.4 Temporal distribution of instances over eight sample categories. The dataset comprises content with high diversity in terms of temporal signatures. . . 119
5.5 Semantic dispersion over time analysis for five sampled images. The y-axis denotes the similarity magnitude where 1 is maximal and -1 is minimal. . . 123

5.6 Temporally bounded cross-modal results (mAP) of DCM-Binned [41] and DCM-Continuous. 127
5.7 Evolution over time for 2 query examples (query timestamps are black-filled). Instances were retrieved from before and after the query timestamp. Image queries were used to retrieve documents through their text. 129
5.8 Evolution over time for 4 query examples (query timestamps are black-filled). Instances were retrieved from before and after the query timestamp. Text queries were used to retrieve documents through their images. 130

List of Tables

2.1 Summary of Multimedia Understanding tasks that can be addressed using mcross-modal embeddings. 23

3.1 mAP performance results across different datasets. The second half of the table concern deep-learning methods. 64
3.2 Comparison between SAM and CMOLRS on the Wikipedia dataset. 65
3.3 mAP results on the NUS-WIDE dataset. 66
3.4 Analysis of the scheduler and f_{mc} impact. 69

4.1 SocialStories dataset information regarding EdFest2016 and TDF2016 events. Seed terms/hashtags, event and crawling time spans are shown. 92
4.2 List of SocialStories categories for EdFest2016 and TDF2016. 93
4.3 Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on NUS-WIDE. . . 97
4.4 Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on EdFest2016. . . 97
4.5 Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on TDF2016. . . . 97

5.1 List of categories of the 20 Years Flickr Images Dataset. 117
5.2 Media Time Period based Inference Results. 121
5.3 Diachronic Semantic Alignment. 125
5.4 Temporal locality bias preservation assessment. 127

Acronyms

BoW	Bag-of-Words.
CCA	Canonical Correlation Analysis.
CNN	Convolutional Neural Network.
DCM	Diachronic Cross-modal (Chapter 5).
IDF	Inverse Document Frequency
ILSVR	ImageNet Large Scale Visual Recognition Challenge.
mAP	Mean Average Precision.
nDCG	Normalized Discounted Cumulative Gain.
NLM	Neural Language Model.
RELU	Rectified Linear Unit.
SAM	Scheduled Adaptive Margins (Chapter 3).
TempXNet	Temporal Cross-modal Embedding (Chapter 4).
UHD	Ultra High-Definition.

Symbols

C	Collection of labelled and multimodal documents (image + text).
c^i	Semantic category or set of semantic categories of a document i.
D_T	Textual original feature space dimension.
D_V	Visual original feature space dimension.
d^i	Multimodal document/instance tuple $(\mathbf{x}_V^i, \mathbf{x}_T^i, ts^i, c^i)$, comprised by an image $\mathbf{x}_V^i$, a textual description $\mathbf{x}_T^i$, a timestamp ts^i and a category label c^i.
$\mathcal{S}$	Embedding space, with dimensionality D, *i.e.* $\mathcal{S} \in D$.
$\mathscr{L}$	Model global loss function.
f_T	Textual projection function that maps an image vector representation to a common embedding space.
TS	Time span interval.
ts^i	Timestamp of a document i.
f_V	Visual projection function that maps an image vector representation to a common embedding space.
$\mathbf{x}_T^i$	Vector representation of a textual description of a document i.
$\mathbf{x}_V^i$	Vector representation of an image of a document i.

Chapter 1

Introduction

1.1 Bridging Vision and Language over Time

Giving computers the ability to fully comprehend an image is one of the main goals of computer vision. This turns out to be a highly relevant problem with major impact in our daily lives.

It is well-known that humans excel at the task of image understanding. Just with a few glances, humans can quickly and accurately understand an image with minimum effort [27, 96]. Among the several aspects responsible for this phenomenon, one is that we easily bring world knowledge (from context, shapes, objects, colors, etc.) to back up and aid our reasoning. Our general visual perception of certain physical world elements (*e.g.* objects, shapes), even though it can be perfected, does not change over time. Namely, after learning how a concept, object, or shape looks like (*e.g.* *sky*, *chair*, *house*, etc.), this knowledge will persist in our visual memory and is used effortless when interpreting and reasoning about images' semantics, together with other contextual elements. The same does not apply to machines, which need to combine both visual and non-visual cues, having distinct heterogeneous computational representations. Moreover, satisfactory object recognition performance is dependent on the availability of large amounts of data.

Apart from objects and shapes, each image has a particular (non-visual) context. In order to fully understand its semantics, one must look beyond its pixels and interpret its **semantic and temporal context**. The context is encoded in descriptions, location, time, and other types of metadata [75]. Descriptions reflect the way humans interpret an image, from a *given instant in time*. They are thus capable of providing additional information, being it either about a visual (e.g. an object present in the image) or non-visual (e.g. event

name) element. Figure 1.1 shows three different images with text descriptions, visually depicting the same semantic concept: wreckage from a natural disaster.

The temporal context of each image plays a key role in defining its semantics. For instance, each image from figure 1.1 has different descriptions, despite representing the same visual content (wreckage), thus having a different temporal context. The rationale is that while in general visual objects do not change over time, and the physics of the real world are static, **textual descriptions are subject to evolution**. Consequently, given a multimodal document d^i, the image $\mathbf{x}_V^i$ with timestamp ts^i, may have an associated text description $\mathbf{x}_T^i$. At a distinct instant ts^j, an image with similar visual content, may be described differently.

(a) Destruction caused by flooding in Pakistan **(Jan-2005)**.

(b) Wreckage of Tsunami in Aceh, Indonesia **(Mar-2005)**.

(c) Destruction of the Earthquake + Tsunami in Japan **(Apr-2011)**.

Figure 1.1: Illustration of *diachronicity* in images: Image depicting the same content, *wreckage*, but with different temporal context, are described differently.

The phenomenon, of textual descriptions evolution, happens when the context of an image changes, due to some external cause, like the occurrence of an event. *When changes in context are evidenced by the temporal dimension, we can say that visual-textual correlations evolved, and the temporal context changed.* Consider the following complementary example. From the perspective of image understanding, when the visual modality is considered alone, images are compared *solely* based on their visual content. The semantics of image contents are estimated from the concepts, scenes and colors, that are identified in each image. Thus, when two visually identical images, from distinct events, are compared, they are deemed as *semantically similar*. However, when the textual description is considered, if the temporal context is different, images are expected to have different textual descriptions. *Models can only differentiate between such two images by considering the temporal dimension.*

This temporal context changes, which in this thesis is framed as the evolution of visual-textual correlations, is grounded on:

a) The occurrence of external events, which have the capability to change the way humans interpret an image [4, 26, 66, 80, 119];

b) Word meaning/usage change across time [40, 41, 136].

The former **a)** occurs on both short and long time spans, being highly dependent on the topic. This is the example of figures 1.1b and 1.1c, which both depict the wreckage left by distinct tsunami events, at distinct time instants, while being *visually* similar. While some events happen at a specific time and place (as defined by McMinn et al. [80]), some events can happen multiple times, and some may not have a location (*e.g.* *Christmas*). In the later **b)**, word meaning change occurs on long time spans, where language itself evolves and certain words are replaced. In practice both aspects **a)** and **b)** are connected, *i.e.* it is likely that changes in the usage of certain words, are triggered by some event.

Regardless of the type of events, given a corpus comprised by content that captures the different events that occurred, the impact of these events in visual-textual correlations is expected to be encoded in the corpus. In other words, the corpus is considered to be a mirror of what happened over the years, thus capturing visual-textual interactions' evolution. This discussion will be further expanded in section 5.4.1.

In the next section we describe the scope and research field in which this thesis is positioned. Namely, we present the starting point scenario in which the thesis hypothesis is built upon (section 1.1.1). Then, we describe the research directions (section 1.1.2) and formalize the scenario and general problem (section 1.1.3).

1.1.1 Cross-modal Embeddings to Bridge Vision and Language

In this thesis, an image semantic and temporal context is defined by its textual description and timestamp. Namely, a textual description, for an image $\mathbf{x}_V^i$ at time instant ts^i, is used to semantically describe the concepts underlying that image. These descriptions usually do not comprise a thorough enumeration of all the objects and shapes on the image, but rather higher-level terms that are responsible for giving context to the image (see Figure 1.1). Therefore, even though remarkable results have been achieved in computer vision tasks – object detection [32, 33, 43, 100, 115, 144, 145], segmentation [43, 72, 73, 74, 89, 108, 143], visual question-answering [3, 76, 104] – these:

a) Mostly focus on understanding images at the object-level, not at the conceptual and contextual level that humans usually refer to images;

b) Use an object recognition approach, which requires large amounts of data to obtain effective detection performance. For comprehensive image understanding and for capturing rich visual-textual correspondences, it is also not feasible to learn a single concept-detector for each possible concept;

c) Discover patterns of interactions between vision and language **under a static world model assumption**.

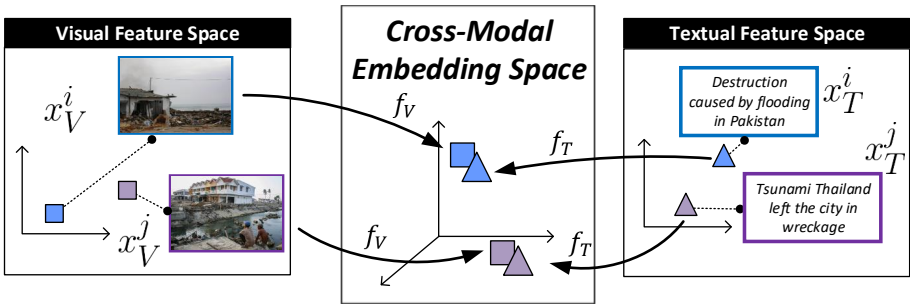


Figure 1.2: Illustration of Cross-Modal Embeddings approach.

An alternative approach for automatic image understanding consists of directly capturing how two modalities, namely vision and language, are correlated. This can be done by **learning how images and texts co-occur over time**.

Cross-modal embeddings, illustrated in Figure 1.2, are an example of such models, which leverage on multimodal machine learning [8, 88] and metric learning [48, 51, 64, 138], to learn an embedding space that structures images and their textual descriptions based on their correlations. Similarly to word embedding models [83, 94], which learn continuous vector representations for individual words, **cross-modal embeddings consist of multimodal (image and text) continuous representations, lying on a space that is common to both modalities, thus bridging heterogeneous representations (*i.e.* from images and texts)**.

Such embeddings allow the learning of rich linear and non-linear correspondences between images and high-level (*e.g.* context-specific vocabulary) concepts by directly capturing patterns of visual-textual interactions from data, instead of learning visual detectors for each potential concept. In other words, visual-textual interactions are represented on a multimodal manifold.

1.1.2 Going Beyond the Static Corpus Assumption

The learning of cross-modal embeddings has been an actively researched topic, but **under the assumption of a static world model** [25, 28, 52, 92, 99, 121, 122, 132, 133]. In other words, the temporal footprint of data interactions and its impact in multimodal correlations has been overlooked. This thesis tackles this gap in the literature, by bringing time into the equation. We investigate models that are **capable of unveiling and representing visual and textual interactions over time**. While there are several ways to represent such interactions over time, we focus on two particular general approaches:

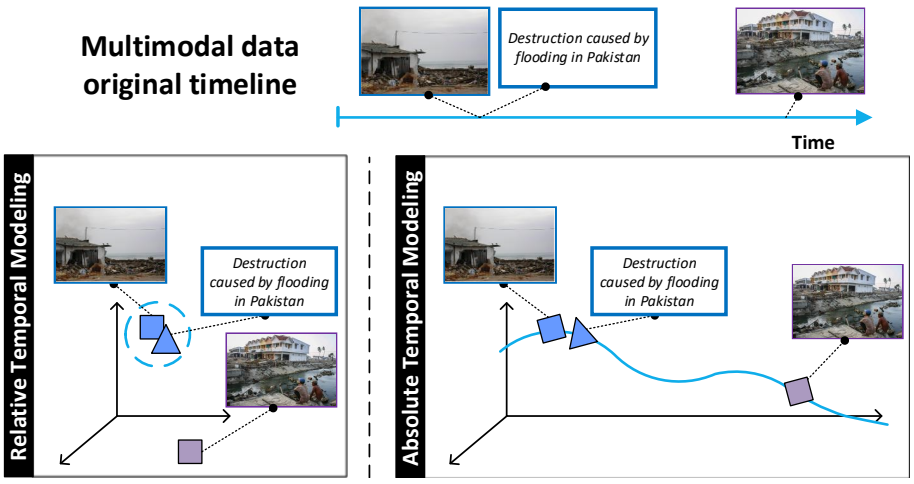


Figure 1.3: Relative vs. Absolute Time dimension Modeling.

a) **Modeling Relative Temporal Correlation** - Pairwise temporal correlations, *i.e.* between two elements (images and/or texts), are unveiled and used to structure data;

b) **Absolute Modeling of the Temporal Dimension** - The temporal dimension is fully preserved, thus enabling unveiling and modeling the evolution of patterns of interaction between visual and textual data.

Each approach, illustrated in Figure 1.3, leads to fundamentally distinct models. In **a)** the temporal dimension is *compressed* and encoded in data neighborhoods. After learning the common embedding space, time is discarded. In **b)** the space explicitly mirrors the original data timeline, with the temporal dimension being fully preserved.

1.1.3 Problem Framework Formal Definition

We start by formalizing the general problem of bridging vision and language, in which the goal is to create a common embedding space that unifies multiple modalities. This corresponds to solving the *heterogeneity gap* [123], in which different correlated modalities, with distinct computational representations must be unified. The goal is to model cross-modal relationships such that one can then compare an element from one modality with all the others, even if one modality is missing.

Without loss of generality, let $C = \{d_i\}_{i=1}^{N}$ be a set of N *visual-textual* instance tuples

$$d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, ts^i, c^i), \tag{1.1}$$

where $\mathbf{x}_V^i \in \mathbb{R}^{D_V}$ and $\mathbf{x}_T^i \in \mathbb{R}^{D_T}$ are the feature representations of the image and textual elements, respectively, ts^i the timestamp and $c^i \in L$ the instances' semantic categories.

L is the set of all semantic categories. Accordingly, D_V and D_T correspond to the image and text original feature representations' dimensionality, respectively. The corpus C has a time span defined by $TS = [t_s, t_f]$, where t_s and t_f are the first and last instants of the corpus, respectively. Throughout this document, let $* \in \{V, T\}$ when referring to *any* of the modalities (visual V or textual T), to avoid notation cluttering.

This thesis focus on models that learn a common, coordinated embedding space $\mathcal{S} \subseteq \mathbb{R}^D$, in which the visual and textual **elements are organized according to their semantic similarity and timestamp**. The space, in its general form, is formally defined by the mappings:

$$\underbrace{f_V(\mathbf{x}_V^i; \boldsymbol{\theta_V}) : \mathbb{R}^{D_V} \mapsto \mathcal{S}}_{\text{Visual Projection}} \qquad \underbrace{f_T(\mathbf{x}_T^i; \boldsymbol{\theta_T}) : \mathbb{R}^{D_T} \mapsto \mathcal{S}}_{\text{Textual Projection}}. \tag{1.2}$$

where f_V and f_T correspond to visual and textual, respectively, *independent* functions, mapping each modality of an instance d^i, to its own D-dimensional embedding space. The two embeddings (one for each modality) are coordinated, in the sense that modalities will be indirectly aligned by enforcing similarity constraints (detailed in section 2.2.3). These are illustrated in Figure 1.2. Both $\boldsymbol{\theta_V}$ and $\boldsymbol{\theta_T}$ correspond to the learnable parameters of the model underlying f_V and f_T. Each function takes as input the original representation of the corresponding modality ($\mathbf{x}_*^i$). From now on, for simplicity we will refer to a common embedding space as two *coordinated* and aligned embedding spaces.

1.1.3.1 Absolute Modeling of the Temporal Dimension

In most approaches *only* original vector representations ($\mathbf{x}_*^i$) are used to map instances. In this thesis we seek for modality projection functions that provide time-dependent embeddings. Therefore, we introduce a continuous model in which functions f_* take both original representations $\mathbf{x}_*^i$ and timestamp information ts^i:

$$\underbrace{f_V(\mathbf{x}_V^i, ts^i; \boldsymbol{\theta_V}) : \mathbb{R}^{D_V} \times TS \mapsto \mathcal{S}}_{\text{Time-Dependent Visual projection}} \qquad \underbrace{f_T(x_T^i, ts^i; \boldsymbol{\theta_T}) : \mathbb{R}^{D_T} \times TS \mapsto \mathcal{S}}_{\text{Time-Dependent Textual Projection}}. \tag{1.3}$$

Similarity between any two visual/textual elements can then be assessed through *cosine similarity*, which corresponds to the magnitude of the angle between their embeddings. The output of f_V and f_T is conveniently normalized such that $\ell_2(f_*(\cdot)) = 1$ and embeddings lie in the unit hypersphere[1]. Then, a similarity function s is defined as:

$$s(\mathbf{x}_*^i, \mathbf{x}_*^j) = f_*(\mathbf{x}_*^i; \boldsymbol{\theta}_*)_* \cdot f_*(\mathbf{x}_*^j; \boldsymbol{\theta}_*)_* \tag{1.4}$$

where · stands for the dot product.

1.1.4 Improving Neural Cross-modal Embedding Learning

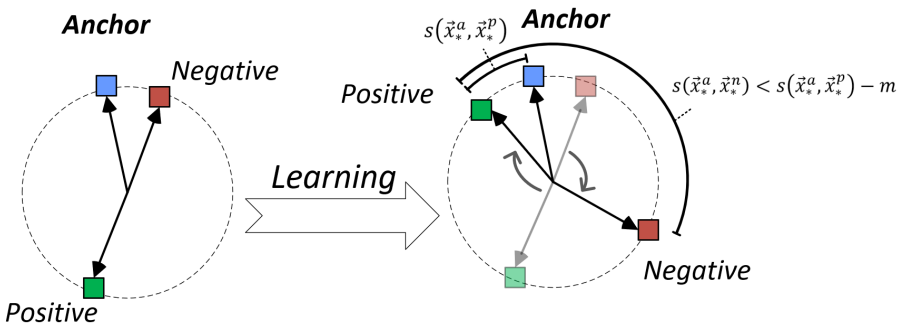


Figure 1.4: Triplet loss illustration. Hinge-loss constraints are enforced over triplets to structure the embedding space.

The projection functions from eq. 1.2 are commonly materialized by a neural network. Among the available loss functions, the *triplet ranking loss* [20, 106] is the most widely adopted loss for neural embedding learning models and metric learning approaches. In the cross-modal embedding learning field, the trend is the same, with state-of-the-art models adopting the triplet loss [121, 124, 135] due to its high expressiveness and effectiveness. In its general formulation, triplets $(\mathbf{x}_*^a, \mathbf{x}_*^p, x_*^n)$, are composed by an anchor element $\mathbf{x}_*^a$ (an image or text), that should be more similar to positive elements $\mathbf{x}_*^p$ sharing a category, than to negative elements $\mathbf{x}_*^n$ not sharing any category, by at least a margin m. Triplet constraints are expressed as

$$s(\mathbf{x}_*^a, \mathbf{x}_*^p) > s(\mathbf{x}_*^a, \mathbf{x}_*^n) + m, \tag{1.5}$$

which are then turned into a differentiable function (based on hinge loss function [46]) This will be detailed in section 2.2.5.1. Figure 1.4 illustrates how triplet loss structures data in a common embedding space. Before learning, the anchor $\mathbf{x}_*^a$ is close to $\mathbf{x}_*^n$ (the negative element) and far apart from the positive element $\mathbf{x}_*^p$. This means that the constraint from eq. 1.5 is violated. After the learning stage, the triplet constraint is enforced such that $s(\mathbf{x}_*^a, \mathbf{x}_*^p) > s(\mathbf{x}_*^a, \mathbf{x}_*^n) + m$. Note that m is just a lower bound on the difference between the two similarities[2].

Throughout this research, we gained deep insights w.r.t. the design and modeling of cross-modal embedding methods. As will be further elaborated in 3, we posit that a

[1]Given that embeddings are ℓ_2 normalized, similarity between projected elements can be efficiently computed through a dot product.

[2]The maximum margin value (upper bound) is 2.

fixed, constant margin m, limits the expressiveness of the loss function in structuring instances on the embedding space. Accordingly, we will investigate how to relax the static margin assumption, and design a triplet-loss formulation that overcomes static triplet-loss limitations and achieves greater expressiveness.

1.2 Objectives and Research Questions

In the previous sections we defined the scope and the domain of this thesis. We provided a brief definition of the problems that are addressed and identified the research fields in which it is positioned. Hereupon, we identify the main objective of this thesis as:

Thesis Main Objective

Investigate neural cross-modal embedding models that model visual-textual interactions over time.

Grounded on this objective, we tackled three distinct research questions that will each be detailed in the next three sections.

1.2.1 Scheduled learning of Adaptive Triplet Loss

State-of-the-art cross-modal embedding learning models adopt the triplet ranking loss formulation which enforces a fixed margin m (see eq. 1.5). It follows that the static formulation has limited flexibility that can compromise the structuring of embedding instances. Therefore, aligned with the problematic described in section 1.1.4, we seek to investigate if by increasing triplet loss expressiveness, and consequently enabling both fine-grain and coarse-grain structuring of multimodal instances in neural cross-modal embedding learning models, one can increase the effectiveness of embedding structure. Thus, aligned with the issues previously described, we formulate our first research question:

Research Question 1 (RQ1)

How to improve neural cross-modal embedding learning models, based on the triplet-loss function, by increasing its expressiveness?

We posit that the static formulation of triplet ranking loss has limited expressiveness w.r.t. performing coarse-grain and fine-grain embedding structuring. Namely, the standard triplet ranking loss does not:

a) Adapt triplet constraints according to the potentially different degrees of similarity, within instances of a triplet;

b) Bound the enforcing of triplet constraints to the model optimization scheme. Namely, neural models are stochastically and iteratively trained. Therefore, triplet constraints should be enforced according to the embedding organization, at each training epoch t.

In Chapter 3, we address **RQ1** and overcome these issues by replacing the static margin formulation, with an adaptive one. Namely, we formulate **an adaptive maximum-margin model** $f_m(\cdot)$, that infers the margin constraints during training. Our model dynamically adapts embedding structuring constraints over triplets, by jointly using semantic similarity and embedding (semantic) category clusters enforcement rules, to obtain an effective embedding organization.

1.2.2 Temporal Cross-modal Structuring

Standard cross-modal embedding models aim at learning a common space which is structured based on semantic (category) information, with the structure reflecting visual-textual correlations. In some scenarios, such as dynamic collections, temporal correlations within instances emerge. This is the case of events that as they unfold, visual and textual correspondences evolve accordingly [26, 119]. Our hypothesis is that incorporating such correlations in the embedding structure, can provide better discriminative power, and therefore result in better embedding structuring. This leads us to formulate our second research question:

Research Question 2 (RQ2)

How can information regarding images and texts temporal correlations be incorporated in cross-modal embeddings? Does this lead to better embedding structuring on dynamic corpora?

Towards answering this question we formulate *Temporal Cross-modal Embeddings*. These aim at enforcing temporal correlations within instances on the embedding space. Therefore, the temporal dimension is *implicitly* encoded in this space, in a **relative manner**: temporal correlations are estimated between each pair of instances (images and/or texts). By being able to structure data not only according to their semantics but also according to their temporal correlations, we obtain a more fine-grain structuring of data that despite being semantically similar, visual-textual correspondences may drift over time.

Chapter 4 addresses **RQ2** by investigating:

a) How can Temporal Cross-modal Embeddings be formulated and materialized;

b) Different models for quantifying relative temporal correlation, and how should estimated correlations be incorporated in the embedding space.

1.2.3 Diachronic Cross-modal Structuring

As discussed in section 1.1, it is natural for visual and textual patterns of interactions to change over time. Therefore, to fully understand an image, both its semantic and temporal context should be captured. This leads to the formulation of our third research question:

Research Question 3 (RQ3)

How to define and learn a diachronic cross-modal embedding space, that bridges vision and language over time, by preserving the temporal dimension, and capturing multimodal interactions' evolution?

To model and *capture the evolution of visual-textual correlations over time*, we seek for representations that preserve the temporal dimension. Data should be seen as a timeline of multimodal instances, instead. Multimodal instances should then be structured in the embedding space over time while **preserving data original timelines**.

Chapter 5 addresses these challenges and aims to answer **RQ3**, by investigating:

a) How should a diachronic cross-modal space be defined, *i.e.* how should images and texts be structured to retain multimodal data evolution, and which properties should be enforced to obtain such structure;

b) Neural architectures that receive time as input (as prior information), and retain the time dimension;

c) Novel diachronic multimedia operations, that enable studying, under different perspectives, the evolution of correspondences between vision and language.

1.3 Contributions and Impact

This research contributes to the ultimate goal of building artificial intelligence models that are capable of bridging vision and language, while modeling their evolution over time. Namely, we exploit neural embeddings to address this problem, and contribute in two distinct and complementary directions:

a) **Chapter 3** - Improving the state-of-the-art in learning cross-modal neural representations, by introducing a scheduled adaptive maximum-margin approach, with

superior expressiveness in capturing and modeling correlations between visual and textual elements;

b) **Chapters 4 and 5** - Bringing temporal information to neural embeddings' structuring, in both a relative (time dimension is discarded) and absolute (time dimension is preserved) manner. Until now, state-of-the-art methods overlooked temporal information, by considering data as a static collection.

1.3.1 Research Publications

The research carried out in this period, resulted in the following main scientific contributions:

1) ACM Multimedia 2019 - Full Paper - In this paper we propose an adaptive formulation for the triplet loss function, and a scheduling training strategy for neural networks that achieves state-of-the-art performance. Materializes the work detailed in Chapter 3:

*D. Semedo and J. Magalhães. **"Cross-Modal Subspace Learning with Scheduled Adaptive Margin Constraints."** In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19. Nice, France: ACM, 2019.*

2) ECIR 2019 - Short Paper - Initial research conducted to validate the importance of temporal information for dynamic corpora, which was demonstrated in the task of keyword extraction (Chapter 4):

*D. Semedo and J. Magalhães. **"Dynamic-Keyword Extraction from Social Media."** In: European Conference on Information Retrieval. ECIR '19. Cologne, Germany: Springer, Advances in Information Retrieval, 2019*

3) ACM Multimedia 2018 - Full Paper - In this paper we propose and evaluate a Temporal Cross-modal Embedding. Materializes the work detailed in Chapter 4:

*D. Semedo and J. Magalhaes. **"Temporal Cross-Media Retrieval with Soft-Smoothing."** In: Proceedings of the 26th ACM International Conference on Multimedia. MM '18. Seoul, Republic of Korea: ACM, 2018.*

4) ACM Multimedia 2019 - Full Paper - Introduces Diachronic Cross-modal Embeddings. Provides a thorough evaluation and demonstration of its novel operations for multimedia understanding. Materializes the work detailed in Chapter 5:

*D. Semedo and J. Magalhães. **"Diachronic Cross-modal Embeddings."** In: Proceedings of the 27th ACM International Conference on Multimedia. MM '19. Nice, France: ACM, 2019.*

1.3.2 Multimodal Research Benchmark Datasets

During the course of this work, and to support the evaluation of the developed models, we created and contributed with two datasets to the scientific community.

***Social Stories*: A Corpus for Social-Media Visual Storytelling** [3] - Comprises data from major real-life events, from the *Twitter*[4] social network. We aimed for events with strong dynamics in terms of temporal variations w.r.t. its semantics, i.e., to the textual vocabulary and visual content. This dataset fills a gap in the research literature, w.r.t. **multimodal annotated collections with highly dynamic content**. This dataset was created to support the research presented in Chapter 4.

All the corpus creation steps from event selection, crawling, filtering and SPAM removal protocols, and initial benchmarks on visual storytelling, are thoroughly detailed in the paper:

*G. Marcelino, D. Semedo, A. Mourão et al. **"A Benchmark of Visual Storytelling in Social Media."** In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. ICMR '19. Ottawa ON, Canada: ACM, 2019.*

This dataset was also the main dataset of the TRECVID 2018 Social-media video storytelling linking (LNK) task [5] [5].

A 20 Years Flickr Images Dataset [6] - We constructed a new large scale weakly-labeled dataset with multimodal instances obtained from the Flickr[7] social network. This dataset fills a gap in the research literature, w.r.t. **long-term** (spanning over 20 years), large multimodal collections. We collect documents related with topics that

[3] https://novasearch.org/trecvid-2018-social-media-video-storytelling-linking/
[4] https://www.twitter.com/
[5] Task web page: http://www-nlpir.nist.gov/projects/tv2018/Tasks/lnk/.
[6] https://novasearch.org/multimodal-diachronic-models/
[7] https://www.flickr.com/

show a dynamic behavior over time such as spike-based and recurring events, covering long time periods. This dataset was created to support the research presented in Chapter 5.

1.3.3 Industry Impact and Use-Cases

The research carried out on this thesis is relevant for use-cases that require automatic capabilities for media understanding. We list below, two use-cases which motivated this work.

1.3.3.1 COGNITUS project: Bringing User-generated Content to Professional Broadcasting

Figure 1.5: H2020 COGNITUS Project - Providing users a "More like being there"Experience, with User Generated Content.

This thesis was partially developed in the context of the COGNITUS H2020[8] project. The goal of this project was to demonstrate the value of user Ultra High-Definition (UHD) contributed content, for the creation of immersive and interactive broadcasts. To this extent, the following use-cases were covered:

- **Engaging the Audience as Camera Crew**: Take advantage of all user cameras present at large scale events (e.g. football match), encouraging users to engage and contribute with UHD content, providing a level of coverage that professional broadcasters cannot emulate.
- **"More like being there"** - Covering geographically-spread events: in such events, professional broadcasters tend to prioritize collecting footage in places with the highest level of audience pull. On the other hand, the public can be at any location of the event and contribute with UHD content that can then be used by broadcasters.

[8]EU H2020 project COGNITUS Grant N° 687640.

In both use-cases, after the content collection stage, professional broadcasters end up with a large collection of related and semantically unstructured visual data, where each visual element has additional metadata like descriptions, timestamps, among others.

The research carried out on this thesis also aimed at addressing the two use-cases above, by enabling the **automatic creation of event-plots**, *i.e.* given an event or a topic of an event, provide the producer a timeline of synchronized user contributed content, covering the different aspects of the event/topic. Additionally, harvesting visual-textual correspondences using the models developed throughout this research, **enables the producer to search and select relevant visual elements** for the creation of a final broadcast for multiple sub-events of a major event, while **aiming for maximal coverage and diversity**.

1.4 Document Organization

The remainder of this document is organized in five chapters, which are briefly summarized below:

Chapter 2 - Related Work: This chapter provides an overview and a critical analysis of the neural cross-modal embeddings literature. It surveys state-of-the-art cross-modal embedding learning methods and approaches that exploit data temporal evidence;

Chapter 3 - Scheduled Adaptive Margin for Neural Cross-Modal Embeddings: This chapter formulates our proposed adaptive maximum-margin approach for representation learning;

Chapter 4 - Temporal Cross-modal Embeddings: This chapter introduces and defines the proposed "Temporal Cross-modal Embeddings", in which temporal correlations are used to guide the learning of a temporal cross-modal embedding;

Chapter 5 Diachronic Cross-modal Embeddings: This chapter formulates the "Diachronic Cross-modal Embedding space", in a model that bridges vision and language over time;

Chapter 6 - Conclusions and Future Work: This final chapter, starts by providing conclusion remarks for this thesis, highlighting the major achievements, discussing limitations to the devised models and proposing future research directions to carry out the line of research started in the thesis.

Chapter 2

Related Work

In this chapter we provide an in depth analysis of the related work and research context in which this thesis is positioned. We start by discussing feature representation for images and texts. Then, we survey in detail the field of cross-modal embedding learning. State-of-the-art works are analyzed in-depth, and their common building blocks are identified. We present and discuss the aspects involved in the optimization. Then we discuss works that successfully leveraged on temporal evidence. We cover approaches that extract temporal information from collections to aid in a specific task. Additionally, we discuss models that capture data evolution. Finally, we present the evaluation metrics commonly used to assess the performance of cross-modal embedding learning methods.

2.1 Feature Representations for Images and Text

In this section we analyze methods for extracting effective image and textual representations for data. It is important to ensure that the features used have a good discriminative power, to enable cross-modal embedding learning models to unveil both coarse-grain and fine-grain correlations, and effectively structure multimodal data in a common embedding space.

2.1.1 Image Representations

Image annotation neural network methods (also referred as deep learning methods [36]) have proved to be highly effective at learning and absorbing discriminative patterns from large collections of data [62]. This trend extends to other fields like face verification [116],

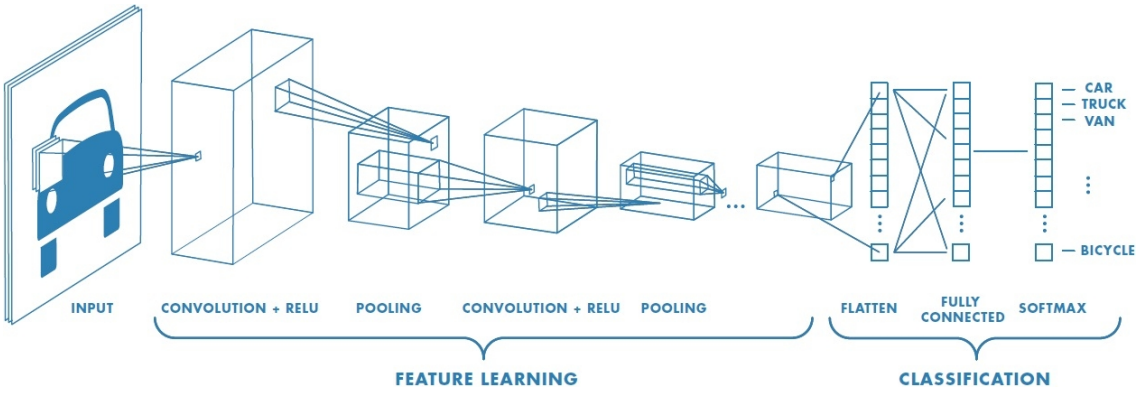


Figure 2.1: Overall Architectural Scheme of Convolutional Neural networks. Adapted from [79]

natural language processing [7], among others. An example of this phenomenon are the results achieved in the ImageNet Large Scale Visual Recognition Challenge [102] (ILSVR), consisting of classifying images with one of 1000 concepts, for which state-of-the-art algorithms, based on neural networks, have surpassed human performance [44, 45].

Among the available deep learning models, Convolutional Neural Networks [36] (CNNs) have been the most widely used for image understanding. CNNs are a type of neural network suitable for images, as they are capable of automatically learn both low and high-level (hierarchical) representations by definition, by learning and applying a sequence of filters, and convolving the image through those filters. The capability of automatically learn filters, directly targeting the task for which the network is trained, potentially eliminates the need for manually selecting low-level and/or high-level features to represent each image. Figure 2.1 illustrates the overall architectural scheme of a CNN.

In general, given an image, these networks apply a sequence of convolutions (with the learnt filters) and pooling operations, at each layer. By definition, CNNs make some assumptions regarding the stationarity of statistics and locality of pixel dependencies, allowing for a reduction in the number of connections and consequently the number of parameters to learn [36, 62]. Through depth and breadth one can control their capacity of identifying high-level data representations and relationships between the input and the output. From this reduction in the number of parameters and with current GPUs processing power, the task of training deep convolutional networks has become feasible. The motivation for building and training deeper CNNs is based on the fact that as the number of layers (depth) of the network increases, so the capacity of detecting more high-level details does, in principle.

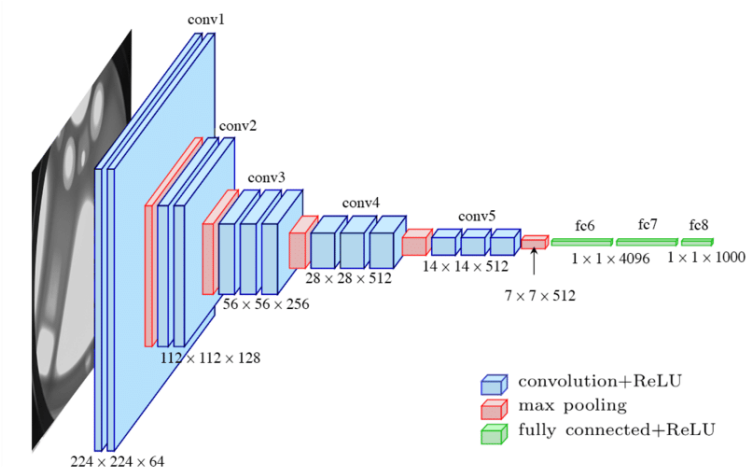


Figure 2.2: VGG-16 Convolutional Neural Network Architecture. Adapted from [30].

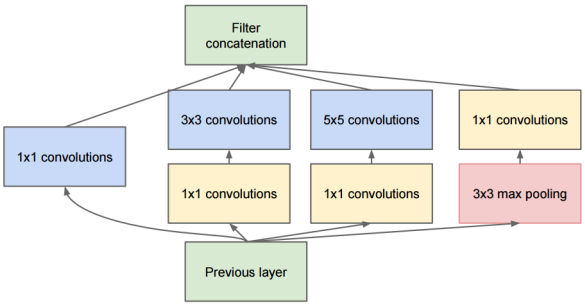


Figure 2.3: Inception module from GoogleNet. Adapted from [115]

CNNs are currently state-of-the-art in image anotation, under a multi-class formulation problem [44, 45, 62, 115]. Among the recent CNN models, AlexNet [62] was the pioneer model, comprising a total of 8 layers, where 5 are convolutional and 3 are dense layers. Instead of traditional non-linearities (e.g. *sigmoid*, *tanh*, etc.) the authors applied the ReLU [87] (Rectified Linear Unit). This activation function, defined as $f(x) = max(0, x)$, is more robust to gradient vanishing problems. The reason is that unlike the sigmoid activation function, which has a range of $[0, 1]$, RELU has a range of $[0, \infty]$, meaning that for large x the gradient will not vanish. Local Normalization is applied on neuron's outputs. The AlexNet model triggered the interest on CNN models for image annotation. Namely, the trend was to design and engineer models with increased depth, while retaining generalization and computational resources required for training.

VGG Net, proposed by Simonyan and Zisserman [109], was designed to be deeper (16 and 19 layers) while keeping simplicity, at the architectural level. Figure 2.2 depicts the architecture of the VGG network with 16 layers. Namely, filters have a fixed size across all layers (3×3), stride and padding (1), and max-pooling layers 2×2. The rationale is that smaller filters (e.g. AlexNet has 11×11 filters in the first layer) require less parameters,

thus allowing for having more convolutions. Following the same rationale of developing a CNN with increased depth, Szegedy et al. [115] proposed the GoogleNet, which has a depth of 22 layers. What essentially sets GoogleNet apart from the previous models, is that the architecture is not based on stacking convolutional layers, but instead on stacking *Inception* modules. These modules (illustrated in Figure. 2.3), instead of performing a sequence of convolutions followed by a pooling operation, perform all the mentioned operations in parallel. To restrain the volume of the output of these operations, 1×1 convolutions are applied before 3×3 and 5×5 convolutions, to factorize the output of the laters, and working as dimensionality reduction technique.

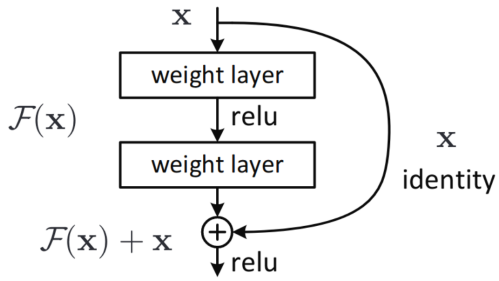


Figure 2.4: Residual learning building block. Adapted from [102].

As the depth of the network increases, it also becomes harder to train, either in terms of avoiding overfitting or due to vanishing/exploding gradients [12, 34]. To overcome this issue, He et al. [44] pursued the idea of letting *residuals* pass through stacked layers, to avoid neurons' saturation. Figure 2.4 depicts a residual block. These residuals are defined as identity mappings x (layer without non-linearities), that are added to the mapping of stacked layers $H(x)$ as $F(x) = H(x) - x$. To restore the original mapping, x is added back again. Using this strategy, the authors were able to develop the ResNet-152, a significantly deeper model comprising 152-layers. The ResNet-152 model won the ILSVRC 2015 [102] competition and even surpassed human performance on the image annotation task. All Across reviewed CNN models, additional techniques are common to avoid overfitting and deal with vanishing/exploding gradients:

1. Dropout [111] - a stochastic regularization technique which essentially consists in randomly dropping neurons and their respective connections during training. The idea is to prevent neurons from co-adapting too much to data, making overfitting less likely;

2. Batch Normalization [53] - Consists of performing mini-batch normalization at intermediate nodes of the network, to cope with changes in layers' input distributions;

3. Data Augmentation: Perform primitive operations to images like translations, horizontal/vertical reflections, and patch cropping, to generate additional samples.

Given the high performance obtained by these methods, one can conclude that they are in fact effective at learning how to semantically discriminate images, and consequently, at learning good feature representations. Accordingly, most cross-modal embedding learning works represent images with features extracted from Convolutional neural networks, pre-trained on ImageNET [102]. The most adopted CNNs are the VGG-19 and the ResNet-50 (Residual CNN with 50 layers). The trend is to extract the penultimate layer of each of these networks. For the VGG-19 this yields a vector of dimension 4096, while for the ResNet-50 it yields a vector of dimension 2048. Lately, the ResNet-50 has become the CNN of election for feature extraction as it not only is more effective, but their representations also have half the number of dimensions, making image understanding systems more efficient.

2.1.2 Text Representations

The way that text is represented computationally is considerable different from images. Textual representations can be divided in two types:

- **Bag-of-Words (BoW) Representations** - Represents texts with the set words comprised in the text. While these representations can preserve multiplicity (*i.e.* terms frequencies, etc.), any information about word ordering and grammar is lost.

- **Distributed vector representations** - Texts are represented using real vectors that encode semantic properties of the text. These can be obtained at the word level (word embeddings [84, 94]) or at the sentence level (sentence embeddings [67]).

In the Bag-of-Words representation, texts are represented as a sparse vector $\mathbf{x}_T \in \mathbb{R}^S$ where S is the size of the dictionary (*i.e.* the number of unique terms). Each dimension of the vector is associated with a term. Then, the dimensions corresponding to the terms that are contained in a text are non-zero (*e.g.* filled with the frequency, term weight, etc.). Figure 2.5 illustrates this representation.

The values of non-zero dimensions, in simple BoW representations often refer to term frequency (*i.e.* how many times does the term appears in the text). However, a common approach is to use weights, that for example reflect term relevance. A widely know approach to find these weights is TF-IDF [78]. In this weighting scheme, a weight

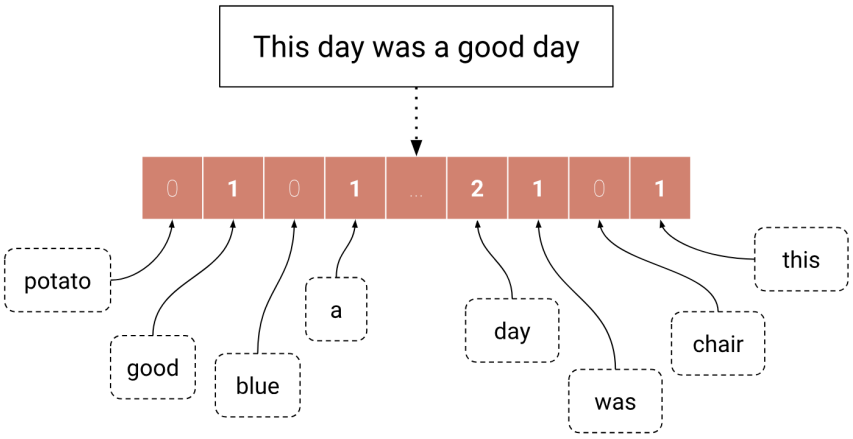


Figure 2.5: Illustration of Bag-of-Word Representation for Sentences. Adapted from [6].

of a word w in a document d is defined as:

$$TF - IDF(w) = TF(w,d) \cdot \underbrace{log\frac{N}{df(w)}}_{IDF} \tag{2.1}$$

where $TF(w,d)$ denotes the frequency of the term in the document d, and $df(w)$ the document frequency, *i.e.* in how many documents from the whole collection, the word w appears. This is a combination of the frequency of the word in d and Inverse Document Frequency (IDF). IDF consists of an estimate of how much discriminative information the word provides, under the rationale that if a word is *rare* across documents, then it should be more important than frequent ones.

An issue of BoW representations is the *curse of dimensionality*, i.e., the number of parameters required to model all the of the words of a sentence, given that realistic vocabularies are very large. *Neural Language models* (NLMs) are language models based on neural networks, that overcome this problem. Bengio et al. [13] proposed to learn distributed words representations (also known as *embeddings*), which are able to represent an exponential number of different words, such that similar words will have close representations (according to a given distance function), allowing for better generalization over the high number of possible sentences in natural language. Word embedding models, such as *word2vec* based on a Skip-gram model [83], *Glove* [94] or the more recent ElMO [95], learn word representations that reflect the context in which individual words appear. They inherit the properties of NLMs w.r.t. distributed representations. These are trained over large corpus, what enables finding rich language regularities.

In practice, it turns out that despite the simplicity of the BoW representation, due to its sparsity property, it can have high discriminative power. Therefore, it ends up being a

Text:
Queuing for the Deep Time Light show: Opening event for Edinburgh Festival 2016

Visual Concepts:
['Castle', 'Sky', 'Clouds', 'Outdoor', 'Vegetation']

Figure 2.6: Example of a multimodal document d^i comprised by an image $\mathbf{x}_V^i$ and a text $\mathbf{x}_T^i$. Illustrates the semantic gap between images and their associated text.

good compromise between semantic expressiveness and efficiency. Namely, state-of-the-art works on cross-modal embedding learning, commonly adopt BoW representation for texts.

2.2 Computationally Bridging Vision and Language

One of the challenges of computer vision, is to develop artificial intelligence models that can effectively bridge vision and language information. Without loss of generality, the goal is to give computers the capability of automatically learn to associate textual information with visual information, and vice-versa.

For humans, this is quite an easy task. While top-performing deep convolutional neural networks classifiers can effectively classify images with a set of 1000 concepts [44, 45, 63, 109, 115], humans are able to recognize a much larger and complex set of concepts. The same applies to text comprehension, where after years of language studying, humans are able to effectively read and comprehend text pieces. One of the reasons is that we easily and unconsciously bring world knowledge (shapes, objects, colors, etc.) to backup our interpretation. The rationale of such procedure may be justified by the fact that both textual and visual content are not drawn from a random distribution but rather each from a *canonical* distribution: same vocabulary with a set of grammatical rules (for text) and human visual perception model (for images) in which the shape of objects is consistent.

These two abilities just described, are key to our cognitive process of reasoning over an image and it's accompanying text, towards identifying what bounds the two modalities. Then, by unconsciously performing a *semantic mapping* of each individual modality to a common, unified, semantic space, we comprehend how a text complements an image, and vice-versa. For example, by identifying the *Castle* in the image from Figure 2.6, and by reading the accompanying text, we immediately conclude that the *Deep Time Show* will be at the *Castle*.

The scientific field of multimodal and cross-modal embedding learning, for visual and

textual data, aims to computationally replicate this process. Conceptually, this problem is framed as given a set of images and their corresponding texts, the goal is to identify patterns of interactions between vision and language, through machine learning, and construct this *common/unified* space. After obtaining this common embedding space, it can be used to address several multimedia understanding tasks. Some of these are listed in table 2.1, where for each task, we represent possible input modalities and corresponding output modalities. Each of the listed tasks are described as follows:

Image Annotation - Annotate an image with one or more tags/keywords;

Image Captioning - Describe an image using natural language, *i.e.* one or more sentences.

Text Illustration - Illustrate a piece of text (sentence, paragraph, among others) with one or more image;

Question Answering - Answer either textual questions or visual question, with an image, a text, or both;

Summarization - Summarize a set of images and/or texts, with a reduced set of images and/or texts;

Multimedia Retrieval - Given a query image or text, retrieve related images and/or texts.

At the core of each of these tasks lies the common need of learning vision and textual correlations, which is at the core idea of cross-modal embeddings.

A challenge that arises when bridging visual and textual data that must be addressed is the **semantic gap**. Figure 2.6 illustrates this issue, where: First, none of the visual concepts are part of the image's text, and second, some high-level concepts, like the *Deep Time Show* and *Edinburgh Festival*, do not have a direct visual materialization. However, by inspecting several multimodal documents of the Edinburgh castle, *during the Edinburgh Festival* (time is crucial to contextualize this example), we discover that the Deep Time Show took place at the Castle. Thus, we bridge modalities (image + text) and discover correspondences between visual elements and textual terms. Cross-modal embeddings tackle both challenges in a principled and versatile manner. Namely, they aim at closing this semantic gap by discovering and analyzing patterns of visual-textual interactions, and representing these patterns in an unified space.

The following section surveys different strategies to learn cross-modal embeddings from multimodal documents comprising images and text.

Table 2.1: Summary of Multimedia Understanding tasks that can be addressed using mcross-modal embeddings.

	Input Modalities			**Output Modalities**		
	V	T	V+T	V	T	V+T
Image Annotation	✓	✗	✗	✗	✓	✗
Image Captioning	✓	✗	✗	✗	✓	✗
Text Illustration	✗	✓	✗	✓	✗	✗
Question Answering	✓	✓	✓	✓	✓	✓
Summarization	✓	✓	✓	✓	✓	✓
Multimedia Retrieval	✓	✓	✓	✓	✓	✓

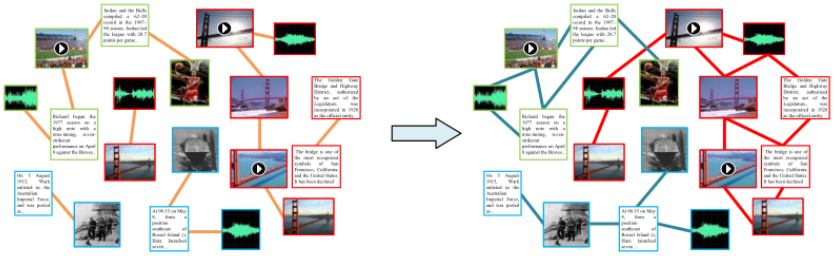
2.2.1 Multimodal Modeling of Images and Text

In this section we analyze works that deal with two modalities, image and text, and leverage on them to tackle one of the tasks presented in the last section.

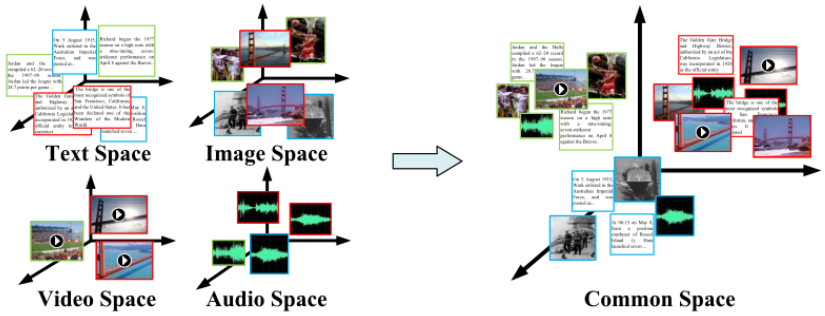
Feng and Lapata [29] proposed a topic modeling approach, representing *visual-textual* pairs using mixtures of latent topics, for image annotation. Then, visual illustration of texts is performed by computing a ranked list of visual terms, w.r.t. to the text, based on the obtained topic model. In fact, given an image and any state-of-the-art image annotation method that outputs words' probability values for given vocabulary, one can construct a similar rank, for each topic (textual) query, and illustrate the topic.

DEVISE [31] departs from the previous work by learning to map images to a textual semantic space. The later is achieved by using a CNN to learn image representations, and then adding an extra layer to the final CNN layer to output the word embedding representation. The resulting network is then optimized using a loss function that enforces similarity between each *visual-textual* pair. Chami et al. [19] used a similar approach to the previous work, in which a non-linear mapping from visual to textual modalities is learned. However, unlike DEVISE, the proposed method projects both modalities, using two distinct projection networks, on a *abstract meta-concept* space. The later consists of a clustered word embedding space, in which each dimension denotes a group concepts. Since in this space pairwise correlated elements are not forced to be aligned, the resulting space is more permissive, allowing for each element to be explicitly aligned with a *set* of other elements.

Yao et al. [135] proposed the Ranking CCA (RCCA) method, which relies on CCA to learn a common sub space for two visual and textual modalities. A key difference is that each pair is enriched with click-trough statistics, i.e. click counts. Thus, after learning the two projections, for each modality, to an aligned space, click-trough information is used to perturb the obtained space when constructing the final rank. The later is achieved by augmenting the objective function with an additional term that captures image-query



(a) Similarity Measurement Approach.



(b) Common Embedding Space Learning Approach.

Figure 2.7: Contrasting between Cross-modal Learning approaches. Adapted from [91].

similarity. Also leveraging on click-trough data, Chen et al. [21] proposes a deep learning approach in which given a query, two attention networks are used to learn to attend to informative parts of each modality (visual and textual). A second attention network, placed on top of the two networks of each modality, is added to learn to better identify which modality is more informative.

While these works develop interesting ways to leverage on visual and textual data, to address some of the tasks listed in table 2.1, they do not truly bridge vision and language. In the next section we will discuss approaches that specifically target this goal.

2.2.2 Cross-modal Embedding Learning

With the aim of computationally bridging vision and language, cross-modal methods exploit latent correlations that are encoded in visual-textual interactions (such as an image and its description). This approach allows the learning of rich linear and non-linear correspondences between images and high-level concepts, by directly capturing patterns of visual-textual interactions from data, instead of learning visual detectors for each potential concept.

Cross-modal methods fall essentially into two categories [91]: Common Embedding Space Learning and Similarity Measurement. The *Similarity Measurement* approach, represented in Figure 2.7a, assess similarity, on data original feature spaces, across different modalities directly, avoiding projecting data to a different space. Instead, a graph is

constructed, where edges between elements of different modalities capture some type of correlation between them (e.g. co-occurrence). The *Common Embedding Space Learning* approach, represented in Figure 2.7a, is based on the rationale presented in the previous section 2.2. Namely, it s rational is that if elements of different modalities are related (*e.g.* appear together in the same document d^i), then there are latent correlations that express such relation, and that can be unveiled/represented in a new space.

This thesis focus on the common cross-modal embedding learning approach, in which the goal is to learn a common embedding space, where latent correlations are unveiled and elements of different modalities are aligned. It is the predominant approach on state-of-the-art methods [28, 92, 121, 133]. The main reason is that methods that adopt this approach, have increased capability of modeling abstractions, that allow the unveiling of complex cross-modal correlations [91]. The use of neural networks for learning data projections, from original features spaces to a target common embedding space, is one of the key elements that enable achieving such capability. As will be discussed later, neural networks can unveil both linear and non-linear correlations, what is crucial to model complex interactions between vision and text.

Formally, the goal is to learn a common embedding space $\mathcal{S} \subseteq \mathbb{R}^D$, in which the visual and textual **elements are organized according to their patterns of interaction, thus bridging heterogeneous representations (*i.e.* from images and texts)**. While this thesis focus on visual and textual modalities, the space is formally defined, and without loss of generality, by two mapping (projection) functions:

$$\underbrace{f_V(\mathbf{x}_V^i; \boldsymbol{\theta_V}) : \mathbb{R}^{D_V} \mapsto \mathcal{S}}_{\text{Visual projection}} \qquad \underbrace{f_T(x_T^i; \boldsymbol{\theta_T}) : \mathbb{R}^{D_T} \mapsto \mathcal{S}}_{\text{Textual projection}} . \tag{2.2}$$

where f_V and f_T correspond to visual and textual, respectively, *independent* functions, mapping each modality of an instance d^i, to its own D-dimensional embedding space. The two embeddings (one for each modality) are constrained to be coordinated and well-aligned. We define $\boldsymbol{\theta_V}$ and $\boldsymbol{\theta_T}$ as the learnable parameters of the model underlying functions f_V and f_T. Figure 2.8 illustrates the application of the two projection functions.

Most approaches that aim at bridging vision and natural language, focus on modeling patterns of interaction from image+text pairs. In this scenario, *only* original vector representations ($\mathbf{x}_*^i$) are used to map instances. While chapters 3 and 4 follow this framework, in chapter 5, we focus on developing approaches that **jointly capture patterns of visual-textual interactions, and their evolution over time**. In practice, we redefine the formal definition of the two projection functions, by augmenting their domain. Namely, we formulate a continuous model in which besides original representations $\mathbf{x}_*^i$, functions f_* also take as input timestamp information.

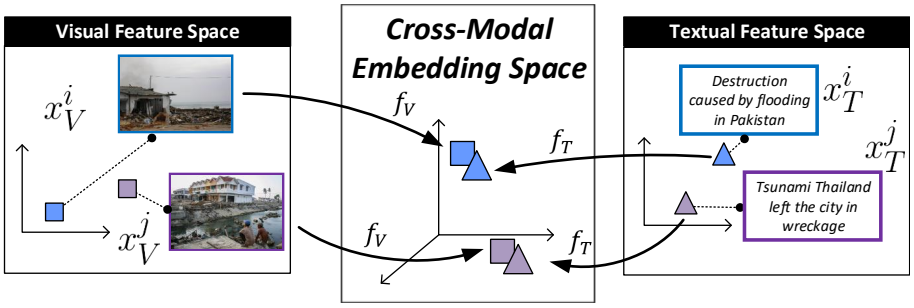


Figure 2.8: Illustration of Cross-modal Embedding Space.

The cross-modal embedding learning formulation supports distinct functionality enabled by a set of data operations. Namely, after the learning stage, the two projection functions are independent. Moreover, each function f_* projects data to a space in which correlated elements, from different modalities, are represented in the same neighborhood, in a modality agnostic space (illustrated Figure 2.8). Jointly, these two aspects enable tackling all the tasks described in the previous section 2.2 and listed in table 2.1. For example, one can directly: a) retrieve data from one modality, given another modality as query (*e.g.* Image Annotation and Text Illustration), b) giving one modality as query, retrieve data from both modalities (*e.g.* Multimedia Retrieval). More complex tasks such as Image Captioning, Question Answering and Summarizaion, may leverage on such projections to navigate over multimodal data, in a common space.

2.2.3 Cross-modal Projection Functions

In this section we survey and discuss methods that materialize the projection functions presented in section 2.2.2.

The rationale is that one wants to learn correlations between images and texts that co-occur and then learn an embedding space that structures images and texts based on those correlations. Given that the goal is to identify patterns of interaction, most methods employ machine learning based approaches.

Works that address the cross-modal embedding learning problem fall in the field of Multimodal Embedding (or Representation) Learning, which is a subfield of multimodal machine learning [8, 88]. Multimodal machine learning aims to research models that can bridge or relate information from multiple modalities (hence the multimodal term). Then, in multimodal embedding learning one seeks to learn embeddings that unify multiple modalities based on how they are related. This are further classified in two distinct types [8]:

a) **Joint representations** - Combines multiple individual modalities original representations, into an unified representation. Formally consist of functions expressed

as:

$$\mathbf{e} = f(\mathbf{x}_{M_1}, \mathbf{x}_{M_2}, \ldots) \tag{2.3}$$

where x_{M_i} corresponds to the original individual representation of modality i and $\mathbf{e}$ a representation that unifies all the modalities.

b) **Coordinated representations** - Process individual modalities independently, but enforce a set of similarity constraints that indirectly align individual modality representations in an unified space. Formally consist of functions expressed as:

$$\mathbf{e}_1 = f_{M_1}(\mathbf{x}_{M_1}) \quad , \quad \mathbf{e}_2 = f_{M_2}(x_{M_2})) \tag{2.4}$$

where similarity between $\mathbf{e_1}$ and $\mathbf{e}_2$ is minimal.

The cross-modal embeddings we seek fall into the category of coordinated representations. In section 2.2.4.1 we will detail the advantages of considering coordinated representations instead of joint representations.

2.2.3.1 Challenges of Learning Cross-modal Embeddings

Solving the task of cross-modal embedding learning comes with a set of challenges. The main problem is the **heterogeneity of multimodal data**. Namely, different modalities, have different computational representations, each with different semantics, that need to be aligned. For example, as discussed in section 2.1.1, images may be represented with low-level features, such as color, textures, contours, among others, or high-level features, obtained from pre-trained neural networks. While all of these are continuous vector representations, they all capture very distinct aspects of images. From the text side, as discussed in section 2.1.2, textual representations are often symbolic (*e.g.* one-hot encoding). Accordingly, cross-modal embedding learning models should be capable of combining data from heterogeneous sources. Moreover, they should be able to model rich correspondences, *i.e.* both simple and complex, between the two modalities.

Then there are also additional problems that are inherent to datasets: how to deal with **incomplete data** and **noise**. The first problem can either be solved by adding more data, which may not be feasible, or bringing additional complementary information to help the structuring of multimodal data. An example of such extra information that can be considered is category information, which leads to supervised or semi-supervised models. In section 2.2.4 we analyze top-performing supervised and unsupervised models, and discuss how some of they incorporate extra information.

While some of this challenges can be implicitly solved by relying on the patterns of visual and textual interactions, which in turn depend on the dataset quality, comprehensiveness and noise, some challenges should be explicitly addressed. Something that is crucial to effectively address the challenge of having heterogeneous representations, is to have specialized projection functions for each modality [25, 28, 88, 121, 133], as defined in section 2.2.2, that focus on mapping from the modality original space, to the common one. Then, even though they are separate, they are learned jointly through the enforcement of a set of constraints (coordinated representations).

2.2.3.2 Early Cross-modal Approaches

We start by discussing early cross-modal linear models that impelled current state-of-the-art works. Namely, in a pioneering work [99], Canonical Correlation Analysis [49] (CCA) was used to learn *linear* projections for each modality in an unsupervised manner. Namely, CCA was used to learn a set of canonical coefficients, that define a subspace where modalities are maximally correlated. Given a set of documents $d^i = (x_V^i, x_T^i)$, image and word vector representations, $\mathbf{x}_V \in \mathbb{R}^{D_V}$ and $\mathbf{x}_T \in \mathbb{R}^{D_T}$, represented as independent random variables $\mathbf{X_V}$ and $\mathbf{X_T}$, respectively, CCA finds canonical representations of the two random variables $\mathbf{X_V}$ and $\mathbf{X_T}$, that maximize their correlation. Let $\mathbf{u} = (\mathbf{x}_V)^T \cdot \mathbf{a_V}$ and $\mathbf{v} = (\mathbf{x}_T)^T \cdot \mathbf{a_T}$. The objective function of CCA is defined as Pearson correlation between $\mathbf{u}$ and $\mathbf{v}$:

$$\rho_{\mathbf{u},\mathbf{v}} = corr(\mathbf{u}, \mathbf{v}) = \frac{E[\mathbf{uv}]}{\sigma_\mathbf{u} \cdot \sigma_\mathbf{v}} = \frac{E[\mathbf{uv}]}{E[\mathbf{u}^2] \cdot E[\mathbf{v}^2]} \tag{2.5}$$

Thus, it learns a set of canonical components $\mathbf{a}_V$ and $\mathbf{a}_T$, consisting of two linear projection matrices (basis), that project original representations onto a new common maximally correlated subspace. The solution of equation 2.5 is obtained by solving the problem as a generalized eigenvalue problem [14].

Representations learned by CCA are *coordinate representations*, *i.e.* it learns two D-dimensional subspaces, where correlation on those subspaces is maximized. Canonical correlations are invariant w.r.t. affine transformations. In other words, CCA is not sensitive to different original feature spaces.

The approach of Rasiwasia et al., based on CCA, was extended in several ways, towards improving its weaknesses (*e.g.* does not capture non-linear correlations) and towards incorporating additional information, such as category information, that can help data structuring (supervised approach).

Gong et al. proposed a multi-view CCA formulation for supervised cross-modal embedding learning, where each document d^i is labeled with a single category. The authors added a third-view to CCA, in which apart from visual and textual modalities, a semantic view is considered, consisting of the category of each document d^i. Then,

a joint space for visual, textual and semantic information is learned, where category information is used to improve data structuring. Ranjan et al. [97] extended CCA for the supervised multi-label scenario, where each document d^i is categorized with one or more categories. In multi-label scenarios, there are implicit many to many relationships between documents. In this approach, the authors use label information to establish correspondences between instances. The main idea is that they break explicit document pairings, *i.e.* image-text pairs that in the vanilla CCA formulation are used to maximize correlation, and stablish new correspondences between image and text elements based on multi-label information.

The main limitation of CCA is that it can only model linear correlations. To address this issue, Zhang and Chen [142] leveraged on a non-linear version of CCA, the Kernel CCA (KCCA), to learn the projections f_V and f_T. Even though it improved performance w.r.t. to the vanilla CCA baseline, it is still limited by the fact that specific (*e.g.* assuming a given distribution) Kernel transformations are applied to data, and thus also fail to learn complex correlations. With the rise of deep learning, projection functions based on neural networks started to be adopted, achieving considerable superior performance. The next section detail neural approaches for cross-modal embedding learning.

2.2.4 Neural Cross-modal Embeddings Learning

For static collections, the task of cross-media retrieval, between visual and textual modalities, has been extensively researched [25, 28, 88, 92, 99, 113, 121, 133].

Neural networks allow the learning of non-linear projections, without the need of committing to specific Kernel transformations, allowing for the unveiling of complex non-linear correlations. Instead, projection weights are learned end-to-end by directly optimizing a given loss function. The works that are discussed in this section all materialize the projection functions f_V and f_T with neural networks. Cross-modal embedding learning methods have proved to be highly effective at learning non-linear projections.

Taking advantage of the later, [28] proposed a Correlation Autoencoder (Corr-AE), which is comprised by two Autoencoders (one for each modality), whose intermediate layers are tied by a similarity measure. Then, the Corr-AE loss function is defined by autoencoder reconstruction error and an additional cost term, measuring the correlation error. Consequently, the network is forced to correctly reconstruct both modalities while learning hidden representations (the projections f_V and f_T) that preserve only common information.

Yan and Mikolajczyk [133] leveraged on Deep Canonical Correlation Analysis [2] (DCCA) to match images and text. DCCA exploits the fact that the CCA objective function can be formulated based on a matrix trace norm. The authors identify overfitting

problems on the original formulation of DCCA. To mitigate this issues, authors modify the original DCCA projections network to incorporate RELU [87] non-linearities (instead of a sigmoid function [2]) and add Dropout after each layer of the text projection network. The modified architecture is shown on Figure 2.11. Peng et al. [93] take a step further by adding extra constraints over inter-modal sample relations, instead of focusing in pairwise *visual-textual* correlation. In [92] the authors model intra and inter modality correlations, to unveil complex modality interactions. To achieve this, image patches were extracted and additional input *visual-textual* pairs, based on these patches were used, allowing for the unveiling of more fine-grained latent correlations. Fan et al. [25], combine image global (CNN features) and descriptive (e.g. caption) representations using a network fusion approach to obtain a richer semantic embedding space. Very recently, Wang et al. [121] proposed to learn a common space using an adversarial learning approach, and achieve state-of-the-art performance. The rationale is to perform minimax game between a feature projector, which is responsible for learning the modality projections, and a modality classifier, which based on the outputs of the feature projector that aim at confusing the modality classifier, will discriminate between each modality (image or text). This will force representations learned by the feature projector to be as modality invariant as possible.

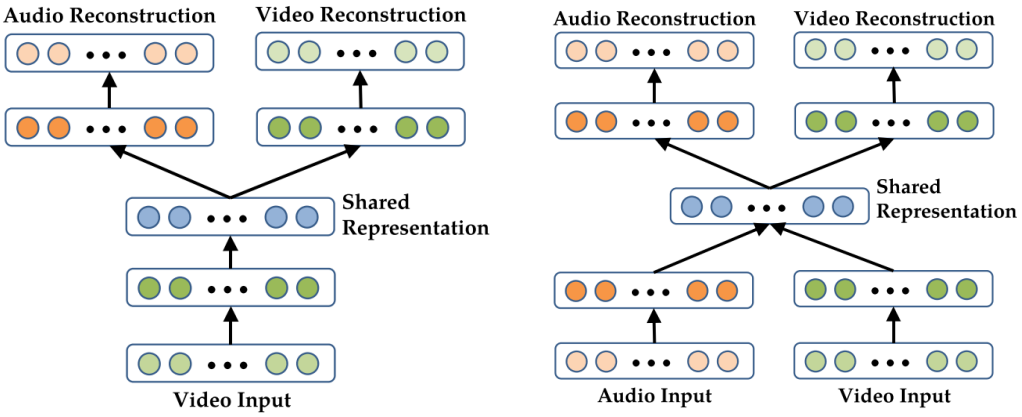
A key element responsible for the effectiveness of neural embedding learning methods, is the loss function. Namely, the rationale is that the loss function should force the outputs of each projection function, to be maximally correlated. For supervised methods, category information if further used to aid the data structuring in the common embedding space. To achieve this, most methods adopt loss functions that employ a *maximum-margin approach*. The idea of such loss functions is to separate, in the common embedding space, not correlated images and text by a given margin. In section 2.2.5 we analyze in detail these loss functions.

Apart from maximizing correlation between different modalities, additional constraints are usually added to the global loss function to impose additional data structuring rules. In [55] center-loss [127] is used to minimize intra-category invariance, under a metric learning approach. The idea of center-loss is to learn an embedding representation for each category (in clustering terms, this would be the category centroid in embedding space), and minimize the distance between embeddings of images and texts, and their corresponding category centroid. A successful approach consists of combining intra-modality semantic category and inter-modality pairwise similarity constraints [92, 121, 124]. Such constraints are commonly enforced over sampled triplets: an anchor image/text, a positive image/text (same category) and a negative image/text (different category). The rationale is to structure the anchor and the positive close to each other, and the anchor far apart from the negative element.

In the following section we will analyze different architectural designs for neural projection functions.

2.2.4.1 Multimodal and Cross-modal Neural Architectures

In this section we will analyze the characteristics of neural architectures used across works, starting with multimodal and then moving to cross-modal approaches.



(a) Multimodal Deep Autoencoder.

(b) Bimodal Deep Autoencoder.

Figure 2.9: Multimodal neural architectures proposed in [88]. Both images were adapted from [88].

Ngiam et al. [88] proposes two types of architectures, based on autoencoders [37], to learn a multimodal embedding for audio and video: Multimodal Deep Autoencoder (Figure 2.9a) and the Bimodal Deep Autoencoder (Figure 2.9b). The first architecture, Figure 2.9a, takes as input a single modality, and is forced to reconstruct both modalities (*i.e.* Audio and Video) solely from a single one. The second architecture, Figure 2.9a, takes as input both modalities, and is forced to reconstruct the same both modalities. In the middle of each of the two architectures is a *shared representation*, which is a layer from which multimodal representations are then extracted after training. The Bimodal Deep Autoencoder outperforms the Multimodal Deep Autoencoder. However, the Multimodal Deep Autoencoder has the advantage that projections for each modality can be made independent, while for the Bimodal Deep Autoencoder, we always need both modalities to project data. A common aspect to both models is that they both use two hidden layers, before the shared representation, both for encoding and for decoding. This enables the model to learn complex transformations to original features, given that data goes through two layers with non-linearities.

The approach of [133] is based on Deep Canonical Correlation Analysis [2]. The original architecture of DCCA is represented in Figure 2.10. The architecture consists of two independent projection networks, one for each modality (referred as view in

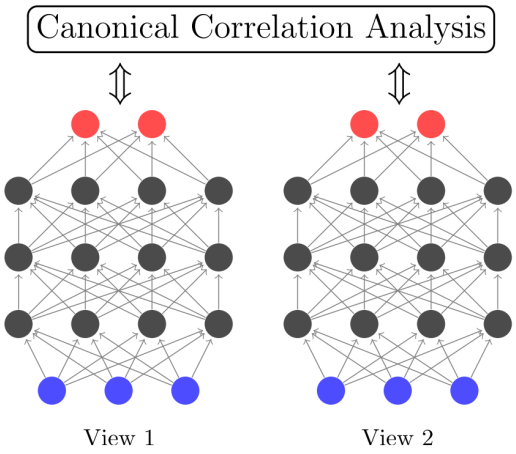


Figure 2.10: Neural architecture of Deep Correlation Canonical Analysis. Adapted from [2].

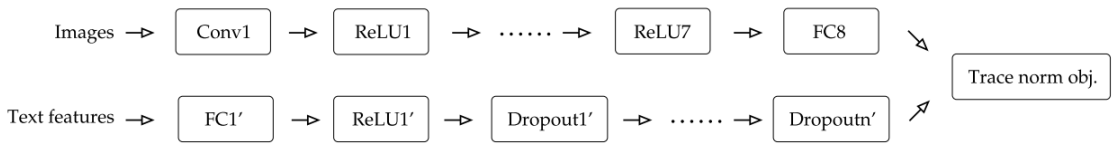


Figure 2.11: Neural architecture of Deep Correlation Canonical Analysis, including the modifications proposed in [133] to deal with overfitting. Adapted from [133].

the paper), that are coupled by the loss function. Namely, the loss function is jointly accommodates the outputs of the two networks. Best performance was achieved with a total of 8 hidden-layers, using a non-saturating sigmoid as non-linearity. In [133], the authors modify the original architecture to make it more robust to overfitting, and to target the modalities used: images and text. Figure 2.11 depicts the modified architecture. There are still two independent networks, but the image projection network is replaced by a convolutional neural network, and RELU non-linearities and dropout is added to layers of the text network.

Similarly to [88], the model of Feng et al. [28] is also based on autoencoders. Namely, two network architectures are proposed: Correspondence Cross-modal Autoencoder (Figure 2.12) and Correspondence Cross-modal Autoencoder (Figure 2.13). The Correspondence Cross-modal Autoencoder, consists of two networks, each corresponding to a vanilla autoencoder. The main difference is that during learning, the two networks are coupled by a code layer represented in Figure 2.12. The coupling is achieved by minimizing the similarity of the outputs of the code layer of each modality network. Then, the functions are optimized by minimizing the reconstruction error. Both architectures use *sigmoid* activation functions. The Correspondence Full-modal Autoencoder, illustrated in Figure 2.13, consists of an extension of the first architecture in the spirit

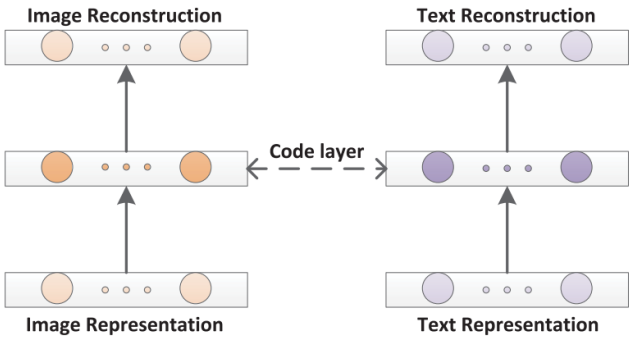


Figure 2.12: Neural architecture of Correspondence Cross-modal Autoencoder. Adapted from [28].

of the Multimodal Deep Architecture from [88], in which the autoencoders from each modality network, are forced to reconstruct not only its own input, but also reconstruct the different modalities. This provides more information about the remaining modality, for each projection network.

In section 2.2.4 we discussed several cross-modal embedding models and now we have discussed their architectures. A common trait to all of these methods, including the state-of-the-art ones, is that all of them use a neural network to materialize projections f_V and f_T. Also common to all these works, is the adoption of two-network (one for each modality) base architecture for projection learning. This means that these works learn *coordinated representations*. Such implementation design ensures that unlike architectures for multimodal embeddings, after training, each projection network can be used independently.

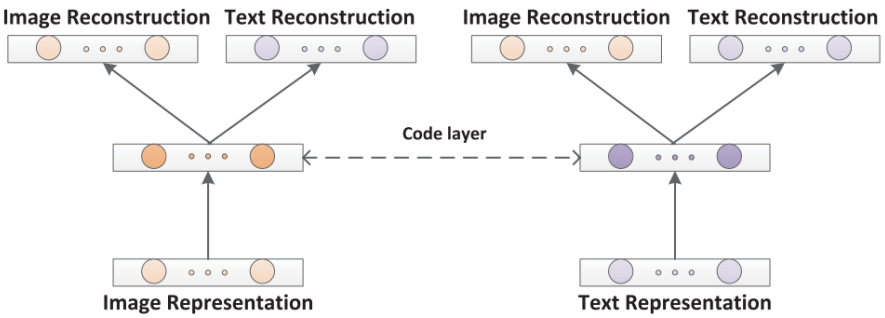


Figure 2.13: Neural architecture of Correspondence Full-modal Autoencoder. Adapted from [28].

2.2.5 Cross-modal Loss Functions and Optimization

Given that in state-of-the-art cross-modal embedding learning approaches, independent projection functions are used, the loss function is generally the core and most important component. While several loss functions have been adopted, most works used some variant of the ranking loss. Among the existing variants, the two that have shown to be more effective are the Contrastive Loss [39] and Triplet Loss [20, 106].

The Contrastive loss was introduced with the aim of mapping pairs of similar input vectors to nearby vectors on an embedding space, and dissimilar ones to distance points. Given a pair of vectors $\mathbf{x}^i$ and $\mathbf{x}^j$, the contrastive loss for the pair is defined as:

$$L(\mathbf{x}^i, \mathbf{x}^j) = \frac{1}{2} \cdot \begin{cases} ||f_*(\mathbf{x}^i) - f_*(\mathbf{x}^j)||_2^2 & , if\ c^i = c^j \\ (max(0, m - ||f_*(\mathbf{x}^i) - f_*(\mathbf{x}^j)||_2))^2 & , if\ c^i \neq c^j \end{cases} \tag{2.6}$$

where $m > 0$ is a margin parameter. Similar pairs contribute to the loss function based on their distance, which should be minimized. Dissimilar pairs (second branch), only contribute to the loss if their distance is greater than m. Then, during training, and at each batch, $\mathbf{x}^i$ and $\mathbf{x}^j$ pairs are created and the contrastive loss from equation 2.6 is evaluated for each of these pairs. In the end, the total loss is the sum of the contrastive losses.

The state-of-the-art approach CCL [92], for cross-modal embedding learning, adopts the contrastive loss. To enforce the contrastive loss it crosses modalities, *i.e.* when the element $\mathbf{x}^i$ is an image, the element $\mathbf{x}^j$ is a vector, and vice-versa.

The triplet loss is more widely adopted among state-of-the-art approaches not only in cross-modal embedding learning [121, 124] but also in other neural embedding learning approaches [106]. Therefore, we dedicate the next section to analyze this loss.

2.2.5.1 Triplet-loss - A Maximum-margin formulation

Modality projections into a cross-modal embedding space must capture both inter-category and inter-modality correlations in that space. To this end, the cross-modal embedding learning problem is commonly formulated using a maximum-margin learning approach, by imposing a set of constraints over pairwise instance's similarity, on the target subspace [92, 121, 124, 135].

In its general formulation, triplets (x_*^a, x_*^p, x_*^n), are composed by an anchor element x_*^a, that should be more similar to positive elements x_*^p sharing a category[1], than to negative elements x_*^n not sharing categories, by at least a margin m. This is formulated as the following constraint:

$$s(x_*^a, x_*^p) > s(x_*^a, x_*^n) + m. \tag{2.7}$$

The constraint above is then enforced over several instances triplets, respecting the triplets' definition. The constraint can then be formulated using angular geometry (cosine similarity) instead of euclidean geometry (distance) to be scale invariant.

Exhaustively creating one constraint for each possible combination of (anchor, positive, negative) would result in a potentially computationally infeasible large set of constraints. In practice, this issue is overcome by stochastically sampling triplets, according to a given triplet sampling strategy [106]. These will be discussed in section 2.2.5.3. Triplet constraints are expressed as $s(x_*^a, x_*^p) > s(x_*^a, x_*^n) + m$, and then turned into a differentiable function, by means of a relaxation under the hinge loss function [46]:

$$\ell(x_*^a, x_*^p, x_*^n; \boldsymbol{\theta}) = [m - s(x_*^a, x_*^p) + s(x_*^a, x_*^n)]_+, \tag{2.8}$$

where m denotes a constant margin, $[x]_+$ the function $max(0, x)$, and $\boldsymbol{\theta}$ is the set of learnable parameters of the model. One of such constraints would then be enforced for each sampled triplet.

For cross-modal embedding learning, methods that adopt the triplet loss impose maximum-margin constraints (eq. 2.8) over the two modality directions ($image \mapsto text$ and $text \mapsto image$), thus simultaneously capturing inter-modality and inter-category correlations [85, 92, 121, 124, 130]. Namely, at every training epoch t, given triplets of the form (x_*^a, x_*^p, x_*^n), where x_*^p and x_*^n stand for positive and negative instances, respectively, w.r.t. an anchor x_*^a, we compute the model loss,

$$\begin{aligned}\mathscr{L}(t, \theta) = &\sum_{p,n} \underbrace{max(0, m - s(x_V^a, x_T^p) + s(x_V^a, x_T^n))}_{image \mapsto text} \; + \\ &\sum_{p,n} \underbrace{max(0, m - s(x_T^a, x_V^p) + s(x_T^a, x_V^n))}_{text \mapsto image},\end{aligned} \tag{2.9}$$

where m denotes the margin and θ the model parameters. This function is evaluated batch-wise. Thus, at each batch, the sampled elements are used to create triplet constraints. This is further discussed in section 2.2.5.3.

2.2.5.2 Triplet Ranking Loss Optimization

Neural embedding learning models are optimized using back-propagation, with mini-batch stochastic gradient descent. In this setting, the loss function $\mathscr{L}$ (eq. 2.9) is evaluated for each batch $\boldsymbol{B} = \{\langle \mathbf{x}_a^1, \mathbf{x}_p^1, \mathbf{x}_n^1 \rangle, \cdots, \langle \mathbf{x}_a^{|\boldsymbol{B}|}, \mathbf{x}_p^{|\boldsymbol{B}|}, \mathbf{x}_n^{|\boldsymbol{B}|} \rangle\}$, comprised by a total of $|\boldsymbol{B}|$ triplets, where the elements of the triplet i of $\boldsymbol{B}$ will be referred as $\mathbf{x}_*^i$.

[1]This formulation is easily extended to the multi-label scenario, where images and texts can have more than one category. In such scenario, positive elements should share *at least* one category.

The general update rule, at step t, is defined as:

$$\theta^{t+1} \leftarrow \theta^t - \eta \cdot \nabla_\theta \mathscr{L}(\boldsymbol{B}, t; \theta^t), \tag{2.10}$$

where the gradient of the function $\mathscr{L}$ is used to update the model weights. We now start by computing the derivative $\mathscr{L}$ w.r.t. the model weights $\boldsymbol{\theta}$. With a static margin, function $\mathscr{L}$, can be written as:

$$\mathscr{L}(\boldsymbol{B}; \boldsymbol{\theta}_*) = \sum_i max(0, m - f_*(x_a^i) \cdot f_*(\mathbf{x}_p^i) + f_*(x_a^i) \cdot f_*(\mathbf{x}_n^i)), \tag{2.11}$$

where the dot product between each two projections is the cosine similarity. Each vector $\mathbf{x}_*^i$ is projected with f_* and ℓ_2 normalized. To simplify the gradient expression derivation, let $s_a^i = \frac{f_{*a}}{||f_{*a}||_2}$, $s_p^i = \frac{f_{*p}}{||f_{*p}||_2}$ and $s_n^i = \frac{f_{*n}}{||f_{*n}||_2}$, with $f_{*a} = f_*(\mathbf{x}_a^i)$, $f_{*p} = f_*(\mathbf{x}_p^i)$ and $f_{*n} = f_*(\mathbf{x}_n^i)$. The *max* function introduces a discontinuity. Therefore, we start by addressing the case in which $m - s_a^i \cdot s_p^i + s_a^i \cdot s_n^i > 0$. Accordingly, for a given triplet i we have that:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \boldsymbol{\theta}} &= \sum_i \Big[\frac{\partial \mathscr{L}}{\partial s_a^i} \cdot \frac{\partial s_a^i}{\partial \boldsymbol{\theta}} + \frac{\partial \mathscr{L}}{\partial s_p^i} \cdot \frac{\partial s_p^i}{\partial \boldsymbol{\theta}} + \frac{\partial \mathscr{L}}{\partial s_n^i} \cdot \frac{\partial s_n^i}{\partial \boldsymbol{\theta}} \Big] \\
&= \sum_i \Big[\frac{\partial \mathscr{L}}{\partial s_a^i} \cdot \frac{\partial s_a^i}{\partial f_{*a}} \cdot \underbrace{\frac{\partial f_{*a}}{\partial \boldsymbol{\theta}}}_{\text{Model specific}} + \frac{\partial \mathscr{L}}{\partial s_p^i} \cdot \frac{\partial s_p^i}{\partial f_{*p}} \cdot \underbrace{\frac{\partial f_{*p}}{\partial \boldsymbol{\theta}}}_{\text{Model specific}} + \frac{\partial \mathscr{L}}{\partial s_n^i} \cdot \frac{\partial s_n^i}{\partial f_{*n}} \cdot \underbrace{\frac{\partial f_{*n}}{\partial \boldsymbol{\theta}}}_{\text{Model specific}} \Big]
\end{aligned} \tag{2.12}$$

All the projections of the triplet elements are obtained from the same two projection functions $f_V(\cdot)$ and $f_T(\cdot)$ (the model), depending on their modality type, *i.e.* image or text. Accordingly, by applying the chain rule, we reach the partial derivative $\frac{\partial f_*}{\partial \boldsymbol{\theta}}$, which corresponds to the derivative of the projection functions w.r.t. to the model weights. As this is model specific (*i.e.* depends on the networks' architecture), we stop applying the chain rule at this point. Expanding equation 2.12 by computing the partial derivatives:

$$\begin{aligned}
\frac{\partial \mathscr{L}}{\partial \boldsymbol{\theta}} &= \sum_i \Big[(-s_p^i + s_n^i) \cdot \frac{\partial s_a^i}{\partial f_{*a}} \cdot \frac{\partial f_{*a}}{\partial \boldsymbol{\theta}} + (-s_a^i) \cdot \frac{\partial s_p^i}{\partial f_{*p}} \cdot \frac{\partial f_{*p}}{\partial \boldsymbol{\theta}} + (s_a^i) \cdot \frac{\partial s_n^i}{\partial f_{*n}} \cdot \frac{\partial f_{*n}}{\partial \boldsymbol{\theta}} \Big] = \\
&= \sum_i \Big[(-s_p^i + s_n^i) \cdot \frac{\partial s_a^i}{\partial f_{*a}} \cdot \frac{\partial f_{*a}}{\partial \boldsymbol{\theta}} + s_a^i \cdot (-\frac{\partial s_p^i}{\partial f_{*p}} \cdot \frac{\partial f_{*p}}{\partial \boldsymbol{\theta}} + \cdot \frac{\partial s_n^i}{\partial f_{*n}} \cdot \frac{\partial f_{*n}}{\partial \boldsymbol{\theta}}) \Big]
\end{aligned} \tag{2.13}$$

The second scenario we have to cover is when $m - s_a^i \cdot s_p^i + s_a^i \cdot s_n^i \leq 0$. When this is the case, we have that:

$$\frac{\partial \mathscr{L}}{\partial \boldsymbol{\theta}} = 0. \tag{2.14}$$

Accordingly, the final update rule expressed in eq. 2.10 is expanded as:

$$\theta^{t+1} \leftarrow \theta^{t} - \eta \cdot \frac{1}{|B|} \sum_{i}^{|B|} \begin{cases} (-s_p^i + s_n^i) \cdot \frac{\partial s_a^i}{\partial f_{*a}} \cdot \frac{\partial f_{*a}}{\partial \theta} + s_a^i \cdot (-\frac{\partial s_p^i}{\partial f_{*p}} \cdot \frac{\partial f_{*p}}{\partial \theta} + \cdot \frac{\partial s_n^i}{\partial f_{*n}} \cdot \frac{\partial f_{*n}}{\partial \theta}) & \text{, if } m - s_a^i \cdot s_p^i + s_a^i \cdot s_n^i > 0 \\ 0 & \text{, otherwise} \end{cases} \tag{2.15}$$

Using this update rule, now each individual modality projection f_V and f_T, with parameters θ_V and θ_T, respectively, can be optimized as follows:

$$\begin{aligned} {\theta_V}^{t+1} &\leftarrow \theta_V^t - \eta \cdot \nabla_{\theta_V^t} \frac{1}{|B|} \left(\mathcal{L}(B; \theta_V^t) \right), \\ {\theta_T}^{t+1} &\leftarrow \theta_T^t - \eta \cdot \nabla_{\theta_T^t} \frac{1}{|B|} \left(\mathcal{L}(B; \theta_T^t) \right). \end{aligned} \tag{2.16}$$

After having the update rules for each modality projection function, it remains to define a strategy to create the batch B of triplets. This is discussed in the following section.

2.2.5.3 Ranking Loss Triplet Mining

In this section we analyze triplet mining strategies, and their adequacy for the task of cross-modal embedding learning. The choice of strategy has proved to be crucial among triplet-loss based methods [106, 124, 129], to achieve convergence. In theory, to approximate the true gradient, given a dataset, one should evaluate the loss $\mathcal{L}$ on all possible triplet combinations of anchor, positive and negative. Such approach can be computationally infeasible and even counter-productive [106]. Instead, the common approach is to create batches B of triplets. There are two approaches to create a batch of triplets B:

Offline Triplet Mining - In this approach, triplets are not created during training, but *offline*;

Online Triplet Mining - In this approach triplets are created during trained, by being sampled directly from mini-batches (when training using mini-batch stochastic gradient descent).

The Offline Triplet Mining approach is considerable less efficient as it requires one extra full pass through the dataset, to sample the triplets. Instead, the Online Triplet Mining, introduced by Schroff et al. [106], is much more efficient as it samples triplets directly from mini-batches. The total number of candidate triplets that can be created from a batch

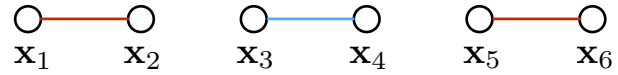
$\boldsymbol{B}$ grows cubically, *i.e.* $|\boldsymbol{B}|^3$ triplets. Additionally, as the network weights are updated after each batch, convergence is faster. While this could be replicated with offline triplet mining, it would require an additional total of $\frac{|C|}{b}$ full passes on the dataset, where $|C|$ is the dataset size and b the mini-batch size. When using the Online Triplet Mining strategy, the batch size plays a crucial role. By increasing the mini-batch size, we increase the number of sampled triplets, thus, the gradients originating from a larger batch will be more *informative*. However, large batch sizes harms the principle of stochastic gradient descent, and has a negative impact in reaching good local optima [128]. Thus, depending on the task and the model, a good balance must be seek.

Another important aspect when creating triplets, is to mine informative triplets [24, 47, 106, 129]. For instance, given a batch of triplets $\boldsymbol{B}$, some triplets will have zero gradient, meaning that the triplet constraint is satisfied. These do not provide useful information for the optimization. One strategy is to mine *hard-negatives* [24, 106]. An hard-negative is defined as the instance $\mathbf{x}_n^i$ such that $\arg\min_{\mathbf{x}_n^i} f_*(\mathbf{x}_a^i) \cdot f_*(\mathbf{x}_n^i)$, *i.e.* its the negative that is more similar to the anchor. However, in [106] the authors observed that this strategy in practice leads to bad local minima early on training, resulting in a collapsed model. There is a trade-off between only considering hard-negatives, and random sampling of negatives. While hard-negatives provide more informative gradients, what may be important specially important near convergence, to avoid stagnation, random sampling provides robustness to outliers [24, 106]. To mitigate some of the issues of hard-negatives, in [106] the authors propose to mine *semi-hard* negatives. These correspond to triplets in which the negative is less similar to the anchor than the positive, *i.e.* $s(\mathbf{x}_a^i, \mathbf{x}_p^i) < s(\mathbf{x}_a^i, \mathbf{x}_n^i)$. Such negatives, lie inside the margin m.

Song et al. [110] proposed an alternative strategy to make full use of the information of each mini-batch, and formulate the Lifted Structured Loss. This is shown in Figure 2.14. Namely, given a mini-batch of size b, in the contrastive loss (eq. 2.6) and triplet loss (eq. 2.8), only $\frac{b}{2}$ and $\frac{b}{3}$, pairs and triplets, respectively, are enforced. The idea is to make use of all possible pairs in the batch. Therefore, given a batch with size b, a total of b^2 triplets are created, where all possible combinations within a batch are considered to create the triplets.

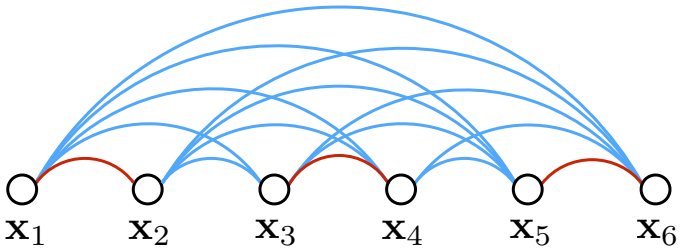
Usually, works that adopt the triplet-loss [18, 92, 106, 121, 130], generate triplets by considering all possible positive elements for an anchor element. In the cross-modal retrieval task, for a given anchor element, every element of an instance of the same semantic category is a positive. However, unlike tasks such as Face Verification and Person re-identification, in which positives refer to images of the same person, content of a semantic category may be potentially more broader.

For the task of cross-modal embedding learning, each batch of size b we have b images and b texts. In our implementations of the triplet loss across the chapters 3, 4



(a) Contrastive Loss. Adapted from [110].

(b) Triplet Loss. Adapted from [110].



(c) Lifted Structured Loss. Adapted from [110].

Figure 2.14: Illustration of different a training batch sampling schemes with mini-batch size of $b = 6$. Red edges and blue edges represent positive and negative instances, respectively.

and 5, we show that it is possible to dramatically simplify the triplet mining process, while still achieving state-of-the-art performance. Specifically, in the spirit of the work of Song et al. [110], we also intend to make full use of batch information. Therefore, our approach consists of considering only as positive candidates of an element i, the modality counterpart of the same instance (*i.e.* $\boldsymbol{x}_V^i$ for $\boldsymbol{x}_T^i$, or vice-versa). This strategy helps capturing *inter-modality* correlations while still enforcing a total of b^2 triplet constraints. Moreover, in order to avoid averaging over zero gradient triplet constraints, we just sum the individual triplet losses (eq. 2.9), and do not divide by the total number of triplet constraints (satisfied and non-satisfied). The work of [61], follows a similar strategy, to learn a multimodal (visual and textual) embedding.

2.2.5.4 Relaxing the Static Maximum-margin Formulation

In state-of-the-art works, this margin is fixed with a constant value for all categories. In fact, this corresponds to a relaxation of the embedding structuring problem, in which the embeddings' semantic similarities are neglected, thus possibly sacrificing optimal data organization. Following this line of reasoning, Li et al. [71] replace the margin by the mean per joint error function, and in [130] the margin is replaced by the correlation of categories in the original feature space. In chapter 3 we depart from methods that adopt a static maximum-margin approach by proposing an adaptive maximum-margin formulation that relaxes the static margin assumption, and infers margin values during training.

2.3 Modeling Temporal Evidence

Accounting for time unveils the temporal dimension in which potentially, discriminative patterns over textual and visual elements will emerge. Across the literature, several works have researched methods to model and incorporate temporal aspects to better understand data. There is enough evidence in the literature that demonstrates how information sources are correlated along a timeline of events, with different media and event types [58, 66, 103, 119, 120].

We refer to temporal evidence as features, signals or representations that can unveil shifts in the patterns of visual and textual data interaction. Despite the literature in this topic being somewhat scarce, some works have proposed techniques to exploit and model such temporal evidence from *visual-textual* data, from whose techniques we can leverage on.

2.3.1 Capturing and Representing Temporal Clues

The literature on modelling and incorporating temporal aspects for multimodal retrieval is very scarce. Uricchio et al. [120] evaluated the value of temporal information, such as tag frequency, for the task of image annotation and retrieval. The authors confirm that some tags reveal a dynamic behaviour that was found to be aligned with Google search trends. This supports our hypothesis regarding the dynamic behaviour of visual and textual correlations on dynamic collections. On the other hand, for orthogonal tasks but directly dealing with social media, time, or more specifically, temporal relevance of elements, has been exploited [66, 103, 119].

Sakaki et al. [103] leveraged on social media posts, and their temporal information, to predict earthquake sizes and directions in real-time. Lau et al. [66] proposed an online trend detection model based on a variant of Online LDA. The Online LDA model is updated in time slices, where each slice corresponds to a documents batch posted on the corresponding time slice. Unlike the original Online LDA, when updating the model with a new time slice, prior counts from previous iterations are removed an a new contribution factor parameter is added to weight the influence of parameter values from previous iterations.

McParlane and Jose [82] were the firsts to explicitly leverage on temporal clues to address the task of image annotation. The authors exploit each individual tag/label temporal distribution to improve the performance on the image annotation task. Their method consists of creating a set of temporal co-occurrence matrices, each with a different granularities (e.g. monthly, weakly, daily, hours of the day, etc.). Then, given a

collection of images previously annotated by some automatic method, the proposed technique consists of removing the annotation with the lowest score, and adding a new one that co-occurs mostly (according to some of the matrices previously defined) with the other annotations. The score is computed as the product of the IDF with the sum of the co-occurrence values. This technique replaces annotations that have few co-occurrences with those that co-occur more. Despite its simplicity, this technique outperformed a baseline which does not account for the temporal dimension. A dataset from a image sharing social platform was used, which has some implications worth discussing. Namely, the authors verified in this dataset that there are tags that are more likely to show temporal patterns, like tags referring to seasonal aspects. These type of tags are the ones which the authors aim to capture. The devised method also applies to our scenario, however, since our data collections are not *topic agnostic*, but instead belong to a story, temporal patterns may be revealed not in seasonal labels but according to their relevance in each period of the story.

In [68] the authors model the temporal dimension using completely different approach. Namely, the authors propose to discover how certain visual elements (e.g. regions), depicting a same object, evolve over time. Here, the temporal evolution refers to the evolution of object shapes and forms over long periods (years). In fact, it corresponds to modeling historical visual style. As opposed to the work previously described in which the temporal dimension of each individual image annotation was exploited according to the timestamp in which each image was posted, here the temporal dimension of each visual element is exploited, instead. The strategy consists of first clustering visually similar image patches with similar labels (date or location). Then, a classifier is trained to identify a same visual element across the whole collection independently of the style (temporal influence). In a last step, for each of the identified groups of visual elements, a style-aware regression model is trained to discriminate subtle stylistic differences between images of a same element. The interestingness and relevance of this approach for our problem lies in the use of anchors (labels) that cluster together groups of similar images, with visual elements that will change over time. This is expected to occur in our setting, when we attempt to model cross-modal correlations evolution.

A pioneer approach was devised by Kim and Xing [58], which formulates a time-sensitive image-retrieval framework, capturing multiple temporal correlations, by a set of temporal attributes that capture correlations between images and its taken date, over different scales (*e.g.* month, year, etc.). Temporal clues were used to improve search relevance at query time, by modeling content streams using a multi-task regression on multivariate point processes. *Visual-textual* occurrences are treated as random events in time and space. Image ranking, under this framework is improved by using a multi-task learning approach that considers multiple image descriptors when capturing temporal

correlations. In the same spirit, Barbieri et al. [10] studied the semantics of emojis change over different seasons. In this work, temporal information was used to improve the effectiveness of emoji prediction models. Similarly, time information is used in event-based media classifiers [75, 117], and has been observed that among images' metadata, time has high discriminative power [75], together with text tags and location.

The hypothesis of chapter 4 presented in section 1.2.2, is directly inspired and supported by the findings of such works, which successfully exploited temporal insights, encoded on dynamic corpora.

2.3.2 Modeling Data Evolution

In this section we review works and approaches that tackle the challenge of modeling data evolution. We start by characterizing the types of data timelines which are considered in this thesis, and contrast them with other types of temporal sequences modeling.

2.3.2.1 Data Timelines

In this section we analyze works that consider a particular scenario in which temporal aspects and cues are encoded in sequences of content, consisting of sequences of images, in which the time span between adjacent images is likely to differ. Thus, temporal aspects are encoded in the *order* or *position* of each individual image in a sequence. When considering sequences, the context is re-defined such that now, images of a sequence not only share the context of the story, but also the context of the sequence to which it belongs. The work of Kim and Xing [59] was a pioneer attempt to exploit storyline graphs to effectively summarize collaborative content from multiple streams (with each social user defining a stream). The devised approach is then evaluated on photo recommendation tasks, that require photo sequence modeling. The authors defined a storyline graph as a directed graph in which each vertex corresponds to a cluster of images, represented by a codeword, with edges connection such clusters, with graphs being time-varying. The time-varying aspect allows edges to vary over a fixed time period (in the paper, periods of a day were considered), originating multiple graphs for each time instant over the considered period. The decision of using image clusters instead of real images is due to scalability issues, since it significantly reduces the number of edges. Graph inference is then applied, with a first step for discovering the topology of the graphs (i.e. which edges are considered) and a second step to estimate the weights of each storyline graph (from each time instant). Edges of graphs of each time instant are discovered by maximizing the log likelihood of each photo stream. Given the learned model, one can then fill incomplete sequences, allowing for storyline based image recommendation. Kim et al. [57] leverage on traveling blogs (sequences of images with associated text written using

a storytelling style) to semantically summarize collections of photo streams, taking into account both visual and textual modalities. The authors tackle the problem by solving two complementary problems: discovery of image-text alignments and photo stream summarization. Both problems are solved under an alternating optimization of two latent ranking SVM framework, consisting of minimizing a regularized margin-based loss, whilst satisfying a set of constraints. This formulation has the ability to exploit similarities between photo streams, such that different sequences can be combined onto a richer sequence. A characteristic of both previous works is that they rely on first-order Markov chains, meaning that only dependences within adjacent sequence elements are captured.

Embedding representations for video, where visual and audio modalities are well aligned at each time instant, has also received a lot of attention [50, 90, 112, 134]. In these approaches, given a visual sequence and its audio, the goal is to learn a common representation that models correlations between modalities and their evolution (*e.g.* motion) over time. These representations are commonly achieved using Sequence Modelling approaches (*e.g.* Recurrent Neural Networks). Such sequences are fundamentally different from data timelines that we find in large collections, spanning several years. They are assumed to show coherence over time and be perfectly aligned (video with audio). In the scenario of this thesis, the concept of a sequence does not exist, as a document d^I is posted once and independently of other instances. **Its temporal evolution is implicitly encoded by the occurrence of semantically similar instances in previous and subsequent time instants.**

2.3.2.2 Capturing Data Evolution Patterns

In order to model the temporal behaviour of data, embeddings must retain temporal correlations [9, 41, 65, 101, 107, 136]. The challenge resides in capturing such correlations and incorporating them in cross-modal embeddings.

Blei and Lafferty [15] proposed a dynamic topic model, Dynamic Latent Dirichlet Allocation (D-LDA), to capture temporal behaviour of data by modeling the evolution of word interactions over time. D-LDA is a dynamic topic modelling (DTM) technique that allows one to analyses the time evolution of latent topics, in documents' collections. Figure 2.15 illustrates the D-LDA architecture. The LDA [16] method, from which D-LDA is based on, represents documents as a finite mixture over a set of estimated latent topic, where each latent topic is characterized by a distribution over words, from which documents are assumed to be generated from. It consists of an *exchangeable* model, as joint probabilities over words are invariant to permutations. D-LDA takes a step further by explicitly addressing topic evolution and dropping the exchangeable property.

Documents are arranged into a set of *time slices* and for each time slice, documents are modelled using a k-component topic model (LDA), where its latent topics at time slice t evolve from latent topics of slice $t-1$.

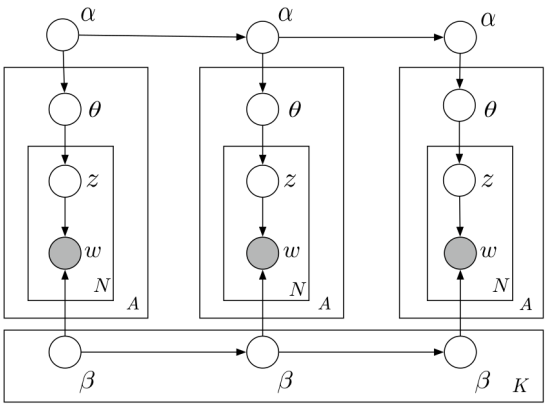


Figure 2.15: Graphical Representation of D-LDA, for three time slices. Each time slice corresponds to an LDA model. Source [15].

A draw-back of topic model approaches is that they treat each word as a symbol, thus they are not continuous and fail to capture semantic similarities between words. Namely, they lack all the properties of distributed representations [13].

Word embedding models aim at learning word representations, such that words that appear in similar contexts are structured close together in the embedding space [83]. *Diachronic Word Embeddings* consist of word embeddings that model word meaning change across time, by encoding words' usage over time [9, 41, 65, 101, 136]. Lately these models have been actively researched to aid the understanding of words' semantic evolution. Figure 2.16 shows one type of analysis that is enabled by diachronic word embeddings: understanding semantic meaning shifts across time.

A common approach to learn such embeddings has been to split text documents into bins (*e.g.* by year), and then train a static Skip-Gram [83] (*word2vec*) model on each bin. Embeddings of adjacent bins are then aligned by learning a linear transformation that performs the best rotational alignment, while preserving cosine similarities [41, 65, 136]. Yao et al. [137] proposed Dynamic Word Embeddings (DWEs) which explicitly address this issue. DWEs can be seen as an extension to continuous word representation models (e.g. word2vec) in which word evolution is modelled. The authors achieved this by partitioning a Point-Wise Mutual Information (PMI) matrix over time slices. The traditional optimization procedure, which envolves factorizing the PMI matrix, is augmented with a term that enforces temporal alignments based on the PMI matrix.

Data binning introduces several issues and limitations. Small bins are required to capture fine-grained interactions, however these may incur in bins with very few data for training. Conversely, with large bins only coarse grained representations can be obtained.

To overcome this, Rosenfeld and Erk [101] recently proposed a continuous approach, in which time is taken as a continuous variable. The model learns an embedding for each word w at each time instant t. The work on chapter 5 to address the hypothesis presented in section 1.2.3 regarding bridging vision and language over time, goes in this direction. However, two aspects invalidate the use of existing word diachronic models: a) unlike words, that are predominant across time instants, each instance is posted only once, invalidating existing alignment strategies, b) in the cross-modal scenario two modalities need to be aligned instead of only one. In Chapter 5 we elaborate on these issues and present a model that overcomes this issues towards learning a diachronic cross-modal embedding.

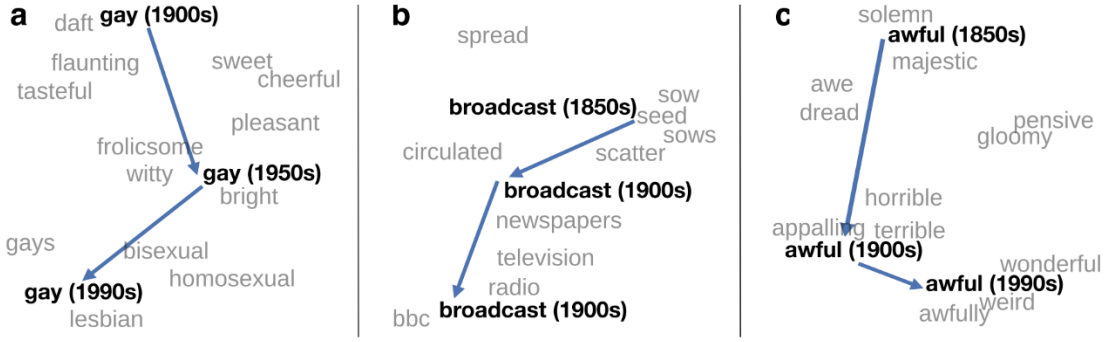


Figure 2.16: Visualization of word shifts across time, based on their similarity with other words under a diachronic word embedding. Source [41].

2.4 Evaluation Metrics

Cross-modal embedding learnings are commonly evaluated on the task of cross-modal retrieval [28, 92, 99, 121, 123, 133]. Namely, two tasks are evaluated: 1) *Image-to-Text* retrieval ($I \mapsto T$) and 2) *Text-to-Image* ($T \mapsto I$) retrieval. Even though the standard evaluation metric is mean Average Precision (*mAP*), we discuss in this section additional metrics that are commonly used.

When measuring the effectiveness of cross-modal embedding models, on the task of cross-modal retrieval, given a query with an Image or a Text, one is interested in obtaining the set of relevant Texts, or Images, respectively. Given a query, an instance is relevant if it shares the same semantic category. Accordingly, the later can be translated to essentially three types of relevance [17]:

- **Binary Relevance** - Each image either is relevant or not;
- **Graded Relevance** - There are multiple levels of relevance (e.g. one image may be more relevant than other, yet both are relevant);

- **Rank Relevance** - results are ordered by similarity and relevance, meaning that results on the top are the most relevant.

For the main task tackled in this thesis, we are mainly interested in Binary relevance. However, for multi-label datasets we are also interested in a combination of Multi-level with Rank Relevance, i.e. we want the most relevant results (the ones with more categories in common) at the top of the rank.

On traditional information retrieval systems evaluation, *Precision* and *Recall* are the most frequently used metrics [17]. Let N denote the size of the results set (e.g. rank size) for a given query, R the number of relevant results, and dr_i the binary relevance of element i.

Precision (P) - Defined as the proportion of relevant documents retrieved by the system:

$$precision = \frac{\sum_{i=1}^{N} dr_i}{N}; \tag{2.17}$$

Recall (R) - Defined as the proportion of relevant documents over the relevant retrieved documents:

$$recall = \frac{\sum_{i=1}^{N} dr_i}{R}; \tag{2.18}$$

Average Precision (AP) - Measures the precision across all recall values. Approximates the area under a precision-recall curve:

$$AveragePrecision = \frac{\sum_{i=1}^{N} dr_i * P@i}{R}, \tag{2.19}$$

Normalized discounted cumulative gain (nDCG) - Evaluates the usefullness (*gain*) of each document based on its position in the rank, while considering graded relevance, i.e. dr_i may have multiple values (e.g. 0, 1, 2, etc.):

$$DCG = dr_1 + \sum_{i=2}^{N} \frac{dr_i}{log_2 i}. \tag{2.20}$$

By sorting the results set by relevance, and then computing the DCG one obtains the ideal DCG, referred as $IDCG$. Finally, the $nDCG$ is defined as the ratio:

$$nDCG = \frac{DCG}{IDCG} \tag{2.21}$$

All the previous metrics can be averaged over all queries.

For ranks, usually recall and precision are evaluated only over the top k documents returned by a query, referred as $P@k$ (precision at k) and $R@K$ (recall at k). These metrics are more suitable to evaluate systems in each we know that the end-user is only interested in the first k results.

The previous metric evaluate the effectiveness of the results of a single rank, for some query. When evaluating a retrieval system effectiveness one should consider and evaluated it under multiple queries. Let Q be the total number of queries. The most widely adopted metric is:

Mean Average Precision (mAP) - Corresponds to the average of AP, over all queries:

$$mAP = \frac{\sum_{i=1}^{Q} AP_i}{Q}; \tag{2.22}$$

Both *mAP* and *nDCG* can also be computed at k, i.e. by considering only the top k positions of the rank.

Chapter 3

Scheduled Adaptive Margin for Neural Cross-Modal Embeddings

To aid the structuring of multimodal spaces, semantic category information is used, to provide supervision to the models. While this achieves better performing models, treating category information solely at a binary level (*i.e.* instance belongs/does not belong to category), is highly strict and too general. Namely, it **assumes that pairwise correlations, within instances of different categories, are all equivalent**. Figure 3.1 (left plot) illustrates this issue, where images of *Sky* are naturally more similar to images of *Mountain* landscapes, than to *Animals*.

In section 1.1.4, we briefly discussed the main ingredients to develop highly effective cross-modal embeddings, and we further expanded the discussion in section 2.2.2. To recap, we now summarize the discussion in the following two aspects:

a) **Neural Projection Functions** - Projection functions f_V and f_T (equation 1.2), materialized by a neural network, are capable of unveiling complex linear and non-linear correlations (discussed in section 2.2.3);

b) **Maximum-margin formulation** - The *triplet ranking loss*, which consists of a variant of the *hinge loss*, is commonly adopted by most state-of-the-art approaches [121, 124] (discussed in section 2.2.5).

We argue that pairwise correlations across instances from distinct categories, may have different levels of correlation. The triplet loss function enforces a set of triplet loss constraints, over sampled triplets (*target instance*; *positive instance*; *negative instance*).

These constraints force the negative to be farther away from the anchor than the positive, based on a fixed margin m (left part of Figure 3.1). Therefore, under a supervised embedding learning setting, triplet loss original formulation assumes that all pairs of instances are equally correlated. If correlation is low, it is okay to separate instances from distinct categories with a large margin m (coarse-grain). Otherwise, if correlation is high, the margin should be lower (fine-grain). This level of expressiveness is not present in the standard formulation of triplet loss.

Accordingly, we develop a **maximum-margin formulation for neural embedding learning**, that is able to account for **both coarse-grain and fine-grain correlations**, between instances, while leveraging on neural networks optimization framework.

Apart from its limited expressiveness, standard triplet-loss function does not account for the iterative and stochastic behavior of neural networks training. Namely, the embedding structure in initial training epochs will still be disorganized due to two reasons:

a) stochastic weight initialization schemes;

b) stochastic mini-batch training.

Therefore, to take the most out of neural-based projection learning, these issues should be addressed. To sum up, *standard triplet-loss does not adapt the constraints imposed by looking at the current subspace organization, (e.g. clusters formed), at each training epoch t*. Then, adding extra terms (*e.g.* smoothing or regularization) to the main loss function, to enforce different types of correlation, possibly at different granularity, may result in contradictory or trade-off optimization objectives, also providing contradictory error information during training. Instead, the different types of correlations that one seeks to capture, should be directly captured in triplet loss constraints. At the same time, the loss function should adapt the constraints imposed, at each training epoch, according to the current subspace structure and enforce semantic clusters formation, *i.e.* promote grouping of instances of the same semantic category.

To overcome these issues, we formulate an adaptive maximum-margin model (SAM), which dynamically adapts embedding structuring constraints over triplets, by jointly using semantic similarity and embedding category clusters enforcement rules to obtain an effective semantic embedding organization. This means that our formulation, **infers triplet-specific margin constraints**, which by enabling the modeling of both coarse-grain and fine-grain interactions, provides higher expressiveness to the model. The right part of Figure 3.1 shows the embedding organization that we aim to achieve, in which instances from different categories are separated according to their correlations (adaptively), instead of all being separated by the same static margin m (left part of Figure 3.1).

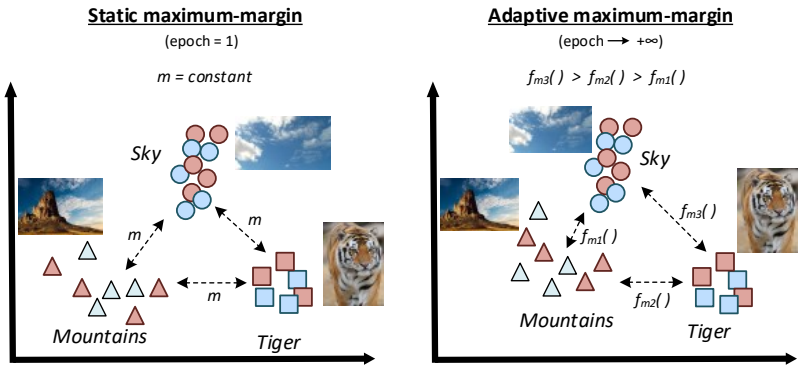


Figure 3.1: Adaptive margin constraints are scheduled to be progressively enforced during the training phase.

In particular our formulation will seek to cover the following two main aspects:

- **Adaptive margin constraints**: we part ways with state-of-the-art methods based on the *triplet-loss* function with a constant margin m between different categories, and introduce a novel adaptive margin function $f_m(\cdot)$ that infers the margin constraints during training;
- **Scheduled activation of adaptive margins**: by considering the incremental learning behaviour of neural architectures [37], we propose a novel *scheduled learning algorithm* that progressively increases the model degrees of freedom to allow a shift from coarse-grain (fixed margin m) to fine-grain (adaptive margins $f_m(\cdot)$) training, as the model reaches a good local optimum. Figure 3.1 illustrates this shift from initial epochs to *epoch t*.

The full adaptive neural maximum-margin approach will now be detailed in the following sections.

3.1 Cross-modal Embedding Space Structure Definition

We start by formalizing the task addressed in this chapter, and the embedding structure that we seek to obtain. Recapping the notation introduced in section 1.1.3, let $C = \{d_i\}_{i=1}^{N}$ be a set of N *visual-textual* instance tuples

$$d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, c^i), \tag{3.1}$$

where $\mathbf{x}_V^i \in \mathbb{R}^{D_V}$ and $\mathbf{x}_T^i \in \mathbb{R}^{D_T}$ are the feature representations of the visual (images) and textual elements, respectively, and $c^i \in L$ the instance (unique) semantic category. Accordingly, D_V and D_T correspond to the image and text features dimensionality, respectively.

In cross-modal embedding learning, the goal is to learn an embedding space in which instances' textual and visual elements, of the same semantic category, will be maximally correlated. The original feature spaces of $\mathbf{x}_V$ and $\mathbf{x}_T$ are dissimilar and cannot be used to perform cross retrieval, as they not only may have different dimensionalty but also encode different characteristics and semantics (heterogeneous representations). To this end, for each original modality space, the goal is to learn the projections:

$$f_V(\cdot;\boldsymbol{\theta}_V) : \mathbb{R}^{D_V} \mapsto \mathbb{R}^D \qquad f_T(\cdot;\boldsymbol{\theta}_T) : \mathbb{R}^{D_T} \mapsto \mathbb{R}^D \tag{3.2}$$

mapping images $\mathbf{x}_V$ and texts $\mathbf{x}_T$ to a common cross-modal embedding, with dimensionality D.

3.1.1 Embedding Properties

We start by stating the fundamental properties that define the structure of a static cross-modal space, and that projections f_V and f_T need to satisfy. The properties are:

- **Property 1**. Two elements will be maximally correlated in the new embedding space (high similarity), *i.e.* projected to the same neighborhood, if they share at one semantic category;
- **Property 2**. Two elements will be minimally correlated in the new embedding space (low similarity), *i.e.* projected onto a distinct neighborhood, if they do not share any semantic category.

In the next section we will detail how these properties are enforced.

3.2 Adaptive Embedding Learning

Modality projections into cross-modal embeddings must capture both inter-category and inter-modality correlations in that space. To this end, the cross-modal embedding learning problem is commonly formulated using a maximum-margin learning approach, by imposing a set of constraints over pairwise instance's similarity, on the target space [92, 107, 121, 124, 135].

For an anchor instance $\mathbf{x}^a_*$, such constraints enforce the similarity between $\mathbf{x}^a_*$ and positive instances $s(\mathbf{x}^a_*, \mathbf{x}^p_*)$, i.e. sharing one category $c^a \in L$, to be higher than the similarity between $\mathbf{x}^a_*$ and negative samples $s(\mathbf{x}^a_*, \mathbf{x}^n_*)$, i.e. not sharing a category, by at least a margin m. This constraint is formulated as:

$$s(\mathbf{x}^a_*, \mathbf{x}^p_*) > s(\mathbf{x}^a_*, \mathbf{x}^n_*) + m. \tag{3.3}$$

The constraint above would then be enforced over each triplet of instances, resulting in a considerable large set of constraints. For training, such constraints are then relaxed using the hinge loss [46], as detailed in section 2.2.5.1. The properties detailed in the previous section 3.1.1 can be jointly enforced through the triplet ranking loss, where given a triplet $(\mathbf{x}_*^a, \mathbf{x}_*^p, \mathbf{x}_*^n)$, an anchor element x_*^a, with semantic category c, is forced to be close to a distinct positive element $\mathbf{x}_*^p$, of the same category $c = c^a = c^p$, but far apart from a negative element $\mathbf{x}_*^n$, *i.e.* from a different category.

3.2.1 Static Maximum-margin Formulation

We start by formulating a loss function $\mathscr{L}$, under this framework, by imposing maximum-margin constraints over the two modality directions ($image \mapsto text$ and $text \mapsto image$), thus simultaneously capturing inter-modality and inter-category correlations. Namely, at every training epoch t, given triplets of the form $(\mathbf{x}_*^a, \mathbf{x}_*^p, \mathbf{x}_*^n)$, where $\mathbf{x}_*^p$ and $\mathbf{x}_*^n$ stand for positive and negative instances, respectively, w.r.t. an anchor $\mathbf{x}_*^a$, we compute the model loss,

$$\begin{aligned}\mathscr{L}(t,\theta) = \sum_{p,n} \underbrace{max(0, m - s(\mathbf{x}_V^a, \mathbf{x}_T^p) + s(\mathbf{x}_V^a, \mathbf{x}_T^n))}_{image \mapsto text} \; + \\ \sum_{p,n} \underbrace{max(0, m - s(\mathbf{x}_T^a, \mathbf{x}_V^p) + s(\mathbf{x}_T^a, \mathbf{x}_V^n)),}_{text \mapsto image}\end{aligned} \tag{3.4}$$

where m denotes the margin and $\boldsymbol{\theta}$ the model parameters. Note that unlike other cross-modal embedding learning works [92, 121, 130], the positive instance x_*^p from each triplet is *only* the opposite modality of the same instance d^i, i.e. $\mathbf{x}_V^p = \mathbf{x}_V^a$ or $\mathbf{x}_T^p = \mathbf{x}_T^a$. This function will be evaluated batch-wise, on a batch of triplet constraints. The sampling strategy is described in section 2.2.5.3.

3.2.2 Limitations of Standard Triplet Loss on Neural Models

When learning an embedding for a given task (metric learning), using neural networks as learnable projection functions that transform input representations to the target embedding space, the summation on equation 3.4 is done over elements of a mini-batch. This has the following implications:

1) Provided that a reasonable number of batch updates are performed, this approach inherits the principles and effectiveness of mini-batch gradient descent [128]. Moreover, as discussed in section 2.2.5.1, it makes training both computationally feasible and efficient, by avoiding enforcing all the possible triplet combinations;

2) Batch gradient updates are based only on triplet constraints from a small set of instances. As a consequence, updates only contain *local* information (at the batch level).

Due to the batch-wise training approach, it is important to enforce triplet constraints that will provide relevant information to update the model. While a common way to achieve this is to design triplet sampling techniques for this purpose (see section 2.2.5.3), using the original triplet loss function will always limit the quality of the information used to update the model. The reason is that the static margin assumption limits the expressiveness of the loss function, by treating all triplets equally.

Stemming from the limited expressiveness of standard triplet loss, two main aspects, that hint how an increase of this expressiveness, may help overcome some embedding learning issues, are now identified.

3.2.2.1 Triplet constraints' set infeasibility

It is possible to have set of constraints that are infeasible, even when enforcing constraints only over a small set (a batch) of instances. This can happen due to the existence of high correlation on feature representations of instances from distinct categories. There may be triplets in which the original feature representations of the anchor and the negative are very similar. This means that even though each element belongs to different semantic categories, they have high correlation based on their original representations. Imposing a large margin on this scenario would provide inaccurate information to the model since we would be asking the model to separate inputs, that in fact are very similar. Using a small margin for these situations, would relax the triplet constraint such that these two instances can be close to each other (*i.e.* separated with a small margin). This somehow resembles the principle of slack variables (soft-margin formulation) for Support Vector Machines [86].

It follows that as discussed, one trivial way to alleviate this problem is to use a small margin. However, from the first condition of the branch equation 2.15, one can see that a small margin has the following implications: **a)** less constraint violations (it is easier to satisfy the constraints), and therefore less updates to the model, and **b)** updates based on triplets with weak violations (the constraint is almost satisfied), which as discussed in section 2.2.5.3 provides weak information to update the model. Therefore, in an ideal scenario, triplet constraints would be made adaptive, in the sense that we allow the margin to be small for some triplets, and large for others.

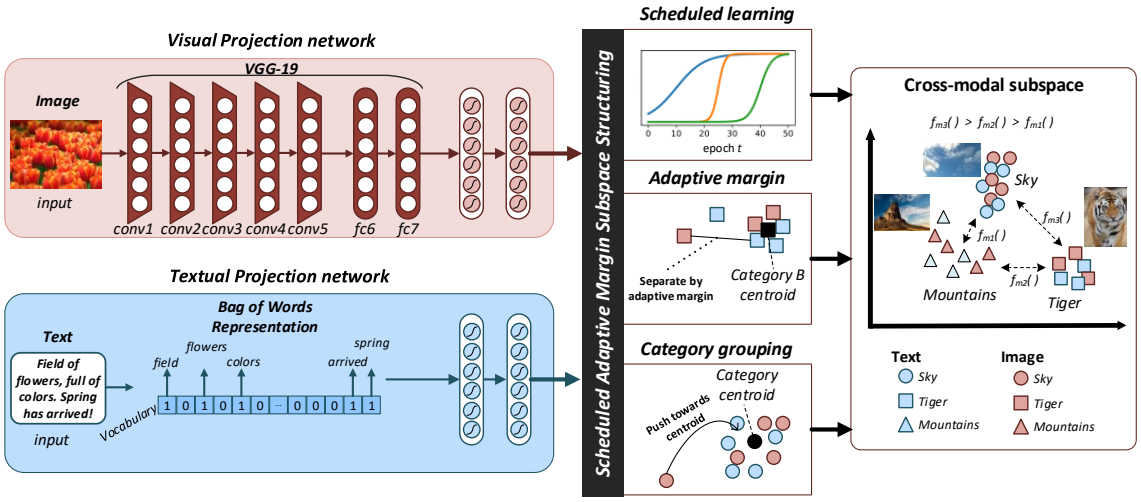


Figure 3.2: SAM model architecture. The model is composed by two sub-networks coupled by the loss function $\mathcal{L}_{SAM}$. At each learning epoch t the loss $\mathcal{L}_{SAM}$ imposes triplet-specific constraints, enforcing cluster formation/preservation and organizing instances according to their semantic similarity.

3.2.3 Adaptive Triplet Loss Formulation

Following the discussion from the previous section, the maximum-margin formulation defined in eq. 3.4 assumes that *any two instances from different categories are equally correlated.* This is reflected by the adoption of a constant margin m. By adapting the margin during training, one can potentially accommodate the issues raised in the previous section: alleviate the triplets' constraints' infeasibility by using smaller margins when adequate, and deal with high correlation between data from distinct categories, by allowing the margin to decrease in this situation. Recent research supports the presented intuition: increasing the expressiveness of the triplet ranking loss leads to better structuring. For instance, Wang et al. [125] propose a Ranked List Loss function which allows performing intra-class structuring, apart from separating data from different semantic category. Instead, we propose to:

1) Incorporate semantic correlations between different categories, into the embedding structuring;

2) Guide the projection learning algorithm, at each epoch, with structure preserving constraints that are derived from the current state of the embedding space.

To achieve this, we design an adaptive margin formulation, defined by a non-negative margin function $f_m(d^a, d^n, t)$, where d^a and d^n correspond to semantically different instances (i.e. belong to different categories) and t denotes the current epoch of the training algorithm. Figure 3.2 illustrates the three components of our formulation.

The margin constraints, for every instance pair, at epoch t, are then reformulated as:

$$s(x_*^a, x_*^p) > s(x_*^a, x_*^n) + f_m(d^a, d^n, t). \tag{3.5}$$

The rationale enclosed in this formulation is that for each pair of instances of *different categories*, $f_m(\cdot)$ outputs a margin that specifies the degree of separation wanted between the considered categories. On every epoch t, the margin will be linked to the pairwise correlation of the instances' original feature vectors and current embedding space structure. Accordingly, the adaptive embedding learning loss function $\mathcal{L}_{SAM}$, at epoch t becomes:

$$\begin{aligned}\mathcal{L}_{SAM}(t, \theta) = & \sum_{p,n} \underbrace{max(0, f_m(d^a, d^n, t) - s(x_V^a, x_T^p) + s(x_V^a, x_T^n))}_{image \mapsto text} \; + \\ & \sum_{p,n} \underbrace{max(0, f_m(d^a, d^n, t) - s(x_T^a, x_V^p) + s(x_T^a, x_V^n))}_{text \mapsto image}.\end{aligned} \tag{3.6}$$

Similar to eq. 3.4, this formulation guides the model towards incorporating semantic information, by sampling the positive and negative elements. Then, we account for the current embedding space organization (at each epoch t), to decide what should be the magnitude of the margin, *i.e.* $f_m(\cdot)$.

This approach resembles the maximum-margin structured SVM [118] formulation which to accommodate complex structured outputs, requires a custom definition of a margin function, that replaces the fixed margin m.

3.3 Scheduled Activation of Adaptive Margins

For neural embedding learning, in the first gradient updates, the space organization is expected to be highly volatile, constantly changing at each epoch. It follows that for neural networks trained using stochastic gradient descent, it is not trivial to estimate beforehand when (i.e. at each epoch) is the model about to reach a local optimum. Thus, we propose an approximation strategy that imposes a hard (*i.e.* a static high magnitude) margin on all triplet constraints on the first few epochs. This allows the model to find an initial coarse organization of the embedding space. Then, as the number of epochs progress, the static constraints give way to triplet specific constraints, that better capture the fine-grain interactions among instances.

3.3.1 Scheduler Function

Inspired by adaptive strategies for neural network training, such as Adam [60], ADADELTA [139] and AdaGrad [23] optimizers, which schedule different learning rates, each using different strategies, we propose a smoothed scheduled shift function from static to an adaptive maximum-margin formulation, as the training algorithm progresses (Figure 3.3). To this end, a scheduled adaptive margin function f_m is defined as:

$$\begin{aligned} f_m(d^a, d^n, t) &= \alpha(t) \cdot f_{am}(d^a, d^n, t) + (1 - \alpha(t)) \cdot m \\ \text{s.t.} \quad \alpha(t) &= \frac{1}{1 + e^{-k \cdot (t - f_a \cdot n_e)}}, \end{aligned} \tag{3.7}$$

where the $\alpha(t)$ is a scheduler function, defined as a *compressed sigmoid*, that gradually activates the adaptive margin, according to the current epoch t. The $\alpha(t)$ function is defined by a smoothing term k, controlling the slope of the function, the total number of epochs n_e and an activation factor $f_a \in [0, 1]$. Figure 3.3 illustrates how each parameter is used to define $\alpha(t)$.

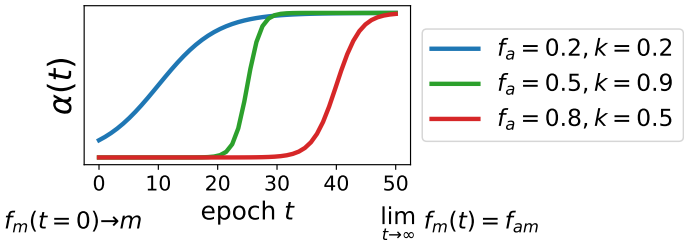


Figure 3.3: Plot of $\alpha(t)$ with $n_e = 50$. The scheduling training enables a smooth transition from static margins to adaptive margins.

3.3.2 Adaptive Margin

In this section we describe how the adaptive margin function $f_{am}(d^a, d^n, t)$ is materialized.

In section 3.2.3, we discussed the two aspects that we aim to capture, towards addressing the issues of standard triplet loss raised in section 3.2.2. The two aspects are briefly summarized as: **1)** flexibility in defining a margin, per triplet constraint, according to instances semantic correlation, **2)** connect the inference of triplet-specific margins to the model optimization, such that margins are inferred while taking into account the current embedding organization.

Accordingly, we formulate f_{am} such that it implements an adaptive margin, encoding: **a)** semantic correlation – estimated from original modality features – between instances from different categories, and **b)** cluster formation enforcement, for each semantic category, according to the epoch t of the algorithm. In particular, we define the adaptive

margin function as

$$f_{am}(d^a, d^n, t) = \lambda \cdot f_{ms}(d^a, d^n) + (1 - \lambda) \cdot f_{mc}(d^a, d^n, t), \tag{3.8}$$

where f_{ms} quantifies semantic correlation, and f_{mc} the similarity between category clusters at epoch t, of two instances d^a and d^n. The parameter λ models the trade-off between the two components. This function now replaces the static margin m.

3.3.2.1 Semantic inter-category pairwise correlations.

From a semantic standpoint, pairwise correlations across categories, will be different (e.g. instances from category *sky* are expected to be more correlated with instances from *clouds* than from *flowers*). In such scenarios, and as discussed in section 3.2.2, inputs may be to similar and imposing a large margin may provide innacurate information to the model. Therefore, in our neural embedding structuring model, the function f_{ms} accounts for such semantic correlations by evaluating similarity on each modality original spaces. The function f_{ms} is then defined as:

$$f_{ms}(d^a, d^n) = \frac{||\mathbf{x}_V^a - \mathbf{x}_V^n||_2 + ||\mathbf{x}_T^a - \mathbf{x}_T^n||_2}{2}. \tag{3.9}$$

From the definition, f_{ms} averages the semantic similarity of both visual and textual modalities, extracted from the modalities' original feature space. The output of this function is normalized to $[0, 1]$.

3.3.2.2 Category cluster formation and preservation.

Given a randomly initialized neural network model (or with a stochastic initialization scheme like Glorot [34], He [44], etc.), the loss can converge to different local optima, thus resulting in different embedding space organization. From this observation, we pose that for near convergence epochs, it is important to restrict model updates, preserving currently formed category clusters and forcing instances to move towards their category cluster, *i.e.* prototype. As a generalization, the prototype of a given category $c \in L$ is defined as the centroid, and is computed as:

$$f_{*-prototype}(c, t) = \frac{1}{|\{\mathbf{x}_*^j : c^j = c\}|} \cdot \sum_{\mathbf{x}_*^k \in \{\mathbf{x}_*^j : c^j = c\}} f_*(\mathbf{x}_*^k; t, \boldsymbol{\theta}_*), \tag{3.10}$$

To materialize the described behavior, we rely on the cosine distance d to define f_{mc} as:

$$f_{mc}(d^a, d^n, t) = \frac{1}{2} \cdot \Big[d(f_{V-prototype}(c^a, t), f_{V-prototype}(c^n, t)) + d(f_{T-prototype}(c^a, t), f_{T-prototype}(c^n, t)) \Big], \tag{3.11}$$

where for a given category c, $f_{V-prototype}(c, t)$ and $f_{T-prototype}(c, t)$ denote the centroid of the visual and textual projections, at epoch t. d stands for the cosine distance $d(\mathbf{x}^1, \mathbf{x}^2) = 1 - s(\mathbf{x}^1, \mathbf{x}^2)$, with s being normalized *a priori* to the $[0, 1]$ range. Essentially, given a pair of instances, f_{mc} evaluates the distance between the corresponding category prototypes, for both visual and textual projections. Grounding the margin on f_{mc} simultaneously enforces cluster formation and preservation. This is achieved since during training, the function f_{mc} will simultaneously attempt to preserve the current embedding space organization and push bad aligned projections towards the corresponding category prototype. To illustrate this, given a triplet constraint in which the category prototypes, of the anchor $\mathbf{x}^a$ and the negative $\mathbf{x}^n$, are close, then f_{mc} imposes a small margin to avoid placing the two instances too far apart from each other, and also far apart from their category prototype. Otherwise, the margin is higher. Specifically, the imposed margin is proportional to the distance between prototypes.

3.3.3 Neural Model and Architecture

To learn projections $f_V(\cdot; \boldsymbol{\theta}_V)$ and $f_T(\cdot; \boldsymbol{\theta}_T)$, following the discussion in related work 2.2.4.1, we consider two independent neural networks, to learn non-linear mappings, as adopted in multiple state-of-the-art works [25, 28, 88, 121, 133]. These are then coupled by a common loss function. Formally, the cross-modal projections are defined as:

$$f_V(\mathbf{x}_V^i; \boldsymbol{\theta}_V) = \underbrace{tanh(\boldsymbol{\theta}_{V_2} \cdot tanh\left(\boldsymbol{\theta}_{V_1} \cdot \mathbf{x}_V^i\right))}_{Visual\ Projection}, \qquad f_T(\mathbf{x}_T^i; \boldsymbol{\theta}_T) = \underbrace{tanh(\boldsymbol{\theta}_{T_2} \cdot tanh\left(\boldsymbol{\theta}_{T_1} \cdot \mathbf{x}_V^i\right))}_{Textual\ Projection}, \tag{3.12}$$

in which $\boldsymbol{\theta}_* = \{\boldsymbol{\theta}_{*_1}, \boldsymbol{\theta}_{*_2}\}$, where $\boldsymbol{\theta}_{*_1}$ and $\boldsymbol{\theta}_{*_2}$ correspond to each modality first and second layers weight matrices, respectively. For each modality, a feed-forward network, comprising 2 fully connected layers is used.

The networks are jointly trained by the common loss function $\mathcal{L}_{SAM}$ (eq. 3.6). For each modality, a feedforward network $f_*(\cdot)$ maps original modality representations onto $\mathcal{S}$ (D-dimensional space), comprising 2 fully connected layers (with dimensions 1024 and D, respectively) and $tanh$ non-linearities.

Algorithm 1 Pseudocode for SAM optimisation.

Initialization: Corpus $\mathcal{C} = \{d^1, \ldots, d^n\}$ of multimodal instances, with $d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, c^i)$;
Initialize network weights: $\boldsymbol{\theta}_V, \boldsymbol{\theta}_T$;
Hyperparameters: λ, k, f_a, embedding space dimensionality D, learning rate η, mini-batch size b;

1: **repeat until convergence**:
2: **for** t epochs **do**
3: Sample mini-batch to create triplets of the form $(\mathbf{x}_V^i, \mathbf{x}_T^i, x_T^n)$ and $(\mathbf{x}_T^i, \mathbf{x}_V^i, \mathbf{x}_V^n)$;
4: Update $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_T$ through back-propagation, with stochastic gradients, using $\alpha(t)$:
5: $\boldsymbol{\theta}_V \leftarrow \boldsymbol{\theta}_V - \eta \cdot \nabla_{\theta_V} \frac{1}{b} (\mathcal{L}_{SAM})$;
6: $\boldsymbol{\theta}_T \leftarrow \boldsymbol{\theta}_T - \eta \cdot \nabla_{\theta_T} \frac{1}{b} (\mathcal{L}_{SAM})$;
7: Update the weight of the adaptive margin:
8: $\alpha(t+1) \leftarrow \frac{1}{1+e^{-k \cdot ((t+1) - fa \cdot ne)}}$;
9: **end for**
10: **return** projection networks, $f_V(\cdot; \boldsymbol{\theta}_V)$ and $f_T(\cdot; \boldsymbol{\theta}_T)$.

3.4 Optimization and Triplet Sampling

We jointly learn both the cross-modal projections $f_{\theta_V}(\cdot)$ and $f_{\theta_T}(\cdot)$, while adaptively performing neural embedding structuring, by minimizing the function:

$$\arg\min_{\boldsymbol{\theta}_V, \boldsymbol{\theta}_T} \mathcal{L}_{SAM}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) \tag{3.13}$$

where $\mathcal{L}_{SAM}$ adaptively organizes instances according to their inter-category and inter-modal correlations. Pseudo-code is illustrated in algorithm 1.

A stochastic sampling strategy is adopted, in which to evaluate $\mathcal{L}_{SAM}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T)$, negative samples are sampled directly from mini-batches. We adopt the strategy of sampling triplets directly from mini-batches, and enforce triplet constraints for all instances, making full use of the information contained in the mini-batch [110]. Specifically, for each instance $\mathbf{x}_T^a$ on a batch, we create triplets between an anchor instance $\mathbf{x}_*^a$ and all the negative instances $\mathbf{x}_*^n$ in the batch. Then, we use as positive element, its modality counterpart, *i.e.* if the anchor is an image ($\mathbf{x}_V^a$), we use a negative text ($\mathbf{x}_T^n$), and if the anchor is a text ($\mathbf{x}_T^a$), we use as negative an image $\mathbf{x}_V^a$. At each epoch, all samples are *seen* by the network. This approach severely reduces the model complexity, while still achieving good local optima. The whole model is then optimized using Stochastic Gradient Descent.

3.5 Evaluation

In this section we evaluate the adaptive maximum-margin formulation, for cross-modal embedding learning. We start by describing the datasets in section 3.5.1, the methodology in section 3.5.2, and finally the training and implementation details in section 3.5.3

3.5.1 Datasets

We evaluate the proposed methods in three widely used cross-modal retrieval benchmark datasets.

- **Wikipedia [99]**. This dataset was made available in the first cross-modal common space embedding learning work [99]. It is comprised by a total of 2,866 *visual-textual* pairs, extracted from Wikipedia's "featured articles", where each article is accompanied by a single image. Each article is annotated with 10 semantic categories. We split the dataset following [28, 93, 99], with 2,173 instances for training, 231 for validation, and 462 for testing.

- **NUS-WIDE [22]**. The NUS-WIDE dataset is comprised by a total of 269,648 instances (images and corresponding tags), from the Flickr network, annotated with one or more categories from a total of 81 distinct semantic categories. For comparison, we follow the protocol of Peng et al. [92]: only instance pairs that belong to a single category are kept and the instances from the 10 categories with more instances[1] are chosen. This results in more than 60,000 instances. Splits are created following [92], resulting in 23,661 instances for testing, 5,000 for validation and the remaining for training.
  NUS-WIDE-10K. Consists of a subset of NUS-WIDE created by strictly following the protocol of [28]: the 10 categories with more instances[1] are chosen, and for each category, 1000 instances are sampled. Only pairs that belong to a single category are considered. Three splits, equally balanced w.r.t. the number of instances per category, are sampled randomly: 8,000 instances for training, 1,000 for validation and 1,000 for testing.

- **Pascal Sentence [98]**. Comprised by 1,000 *visual-textual* pairs, from the 2008 PASCAL development kit, categorized within 20 categories, with 50 instances from each category. We follow [28, 93] and randomly split the dataset with 800 instances for training, 100 for validation and 100 for testing, while keeping the same number of instances per category, in each split.

[1] Top-10 categories: 'person', 'animal', 'sky', 'window', 'water', 'flowers', 'food', 'toy', 'grass', 'clouds'.

3.5.2 Methodology

We evaluate the retrieval performance using mean Average Precision (*mAP*), which is the standard evaluation metric for cross-modal retrieval [28, 92, 99, 121, 123, 133]. We follow [54, 92, 99, 141] and compute *mAP* for *all the retrieved results*. For *mAP*, an instance is relevant if it has the same category. Two tasks are evaluated: 1) *Image-to-Text* retrieval ($I \mapsto T$) and 2) *Text-to-Image* ($T \mapsto I$) retrieval. Core parameters of SAM are analysed to assess their impact in the performance. Each *mAP* result reported of our method corresponds to the average of 5 runs. We complement our evaluation with a qualitative analysis.

We compare our proposed approach, SAM, with a total of 11 state-of-the-art works, on the task of cross-modal retrieval. Namely, we compare against:

- CCA [49] - Canonical Correlation Analysis, a linear embedding learning approach;
- CFA [69] - Cross-modal Factor Analysis. Based on Latent Semantic Indexing (LSI) but extended to support off-line supervised training;
- KCCA [42] - Kernel version of CCA;
- Corr-AE [28] - A deep Correspondence Autoencoder;
- JRL [141] - Graph-based approach that learns a common space using category information, with semi-supervised regularization and sparse regularization;
- LGCFL [54] - Supervised approach that considers unpaired data;
- DCCA [133] - Deep Canonical Correlation Analysis, a neural network-based extension to the CCA algorithm;
- CMDN [93] - Neural network-based approach that jointly models intra-modal and inter-modal information;
- Deep-SM [126] - Deep semantic matching model relying on a fine-tuned CNN;
- ACMR [121] - Learns a common embedding space using an adversarial learning approach;
- CCL [92] - Model intra and inter-modality fine-grain correlations by not only using the original image but also by extracting image patches.

All the baselines are described in 2.2.2.

3.5.3 Training and Implementation Details

Networks are jointly trained using Stochastic Gradient Descent, with 0.9 Nesterov Momentum, and a learning rate $\eta = 5 \times 10^{-3}$, with a decay of 1×10^{-6}. The model with lowest validation error is kept. Mini-batch size is set to 200 for all datasets, and the total number of epochs is set to 100. The margin m is set to 1.0.

Figure 3.2 depicts the full architecture. For each neuron, *tanh* non-linearities are applied. Dropout with $p = 0.1$ is applied to the first hidden layer. For semantically rich image representations, we extract features from a pre-trained convolutional neural network on the task of image classification. Namely, we use a pre-trained VGG-19 [109], with the last fully connected layer removed (softmax) to extract features. Following [28], for texts, we adopt a BoW representation, with 1000-D vocabulary size for NUS-WIDE-10k and Pascal Sentences, and 3000-D for Wikipedia.

3.6 Results and Discussion

3.6.1 Cross-modal Retrieval

In this section we evaluate SAM against state-of-the-art methods (listed in section 3.5.2), on the task of cross-modal retrieval, on four different benchmark datasets.

3.6.1.1 Pascal Sentences dataset

Table 3.1 shows the results obtained. Our method outperforms all the compared methods, on both $I \mapsto T$ and $T \mapsto I$ settings. Namely, SAM achieved a relative improvement of $\approx$ 12.5%, w.r.t. the second best performing method, CCL, on the average of $T \mapsto I$ and $I \mapsto T$. CCL models intra-modality and inter-modality correlations through distinct constraints, using a strategy that balances both types of correlation constraints. These are then superseded by a ranking loss function in which a static margin is used. Instead, SAM adopts an adaptive margin formulation, in which intra and inter modality correlations are directly modeled in a single constraint. The best result was achieved with $\lambda = 0.25$, $f_a = 0.4$ and $k = 0.1$, meaning that SAM started smoothly activating the adaptive margin at about half the training epochs, revealing preference for starting using f_{am} sooner. The semantic similarity component f_{ms} slightly contributes to the effectiveness of the model. Notwithstanding, the component f_{mc} has revealed to be the most important one (75%), effectively guiding the embedding space structuring.

Table 3.1: mAP performance results across different datasets. The second half of the table concern deep-learning methods.

Method	Pascal Sentences			NUS-WIDE-10k			Wikipedia		
	$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg
CCA [49]	0.203	0.208	0.206	0.167	0.181	0.174	0.298	0.273	0.286
CFA [69]	0.476	0.470	0.473	0.406	0.435	0.421	0.319	0.316	0.318
KCCA [42]	0.488	0.446	0.467	0.351	0.356	0.354	0.438	0.389	0.414
LGCFL [54]	0.539	0.503	0.521	0.453	0.485	0.469	0.466	0.431	0.449
JRL [141]	0.563	0.505	0.534	0.466	0.499	0.483	0.479	0.428	0.454
Corr-AE [28]	0.532	0.521	0.527	0.441	0.494	0.468	0.442	0.429	0.436
DCCA [133]	0.568	0.509	0.539	0.452	0.465	0.459	0.445	0.399	0.422
CMDN [93]	0.544	0.526	0.535	0.492	0.542	0.517	0.487	0.427	0.457
Deep-SM [126]	0.560	0.539	0.550	0.497	0.478	0.488	0.478	0.422	0.450
ACMR [121]	0.538	0.544	0.541	0.519	0.542	0.531	0.468	0.412	0.440
CCL [92]	0.576	0.561	0.569	0.481	0.520	0.501	0.505	**0.457**	0.481
SAM	**0.637**	**0.643**	**0.640**	**0.563**	**0.594**	**0.579**	**0.518**	**0.457**	**0.487**

3.6.1.2 NUS-WIDE-10k dataset

From the results on table 3.1, we can see that our method also achieved the best performance when compared to all methods, on both cross-modal retrieval directions. It outperformed both traditional cross-media models (top rows of table 3.1) and the most recent deep learning methods. w.r.t. the second best performing method, ACMR, which uses an adversarial approach for embedding learning, we obtain a relative improvement of $\approx 9\%$, on the average of $T \mapsto I$ and $I \mapsto T$. This confirms the importance of moving towards an adaptive margin formulation. The best result was obtained with $\lambda = 0.05$, $f_a = 0.9$ and $k = 0.1$. Hence, in contrast to the results on the Pascal sentences dataset, the method started activating the adaptive margin near the last epochs of training. We believe this is due to the fact that as the dataset is larger, more constraints with a large margin need to be enforced, until a good coarse-grain structuring of the embedding space is achieved, to then start enforcing more fine-grain triplet constraints. Moreover, once again, more importance was given to the cluster enforcement and preservation (95% of the weight). Our method obtains a high *mAP* on both directions, but performs better on the $T \mapsto I$ direction, following the tendency of all other methods. We believe that the reason is that visually, some categories have very similar content (e.g. *sky* vs. *clouds*). However, the text in this dataset correspond to tags, which due to the sparsity of BoW representation, turns out to have good discriminative properties.

Table 3.2: Comparison between SAM and CMOLRS on the Wikipedia dataset.

Methods (t-*mAP*@100)	Avg.
CMOLRS [130]	0.413
SAM	**0.541**

3.6.1.3 Wikipedia dataset

As with the previous datasets, our method outperforms all the compared methods. It happens that on the Wikipedia dataset, categories are very broad (e.g. Art & Architecture, Media, etc.), with texts and images of the same category being highly diverse. Therefore, in this dataset, given the small amount of instances available for training, it is harder to align modalities. As this is reflected in original feature representations, the function f_{ms}, which organises instances according to semantic similarity on original features, ends up not helping structuring the space. Supporting this observation is the fact that the best result was obtained with $\lambda = 0.05$. The category cluster formation and preservation, enforced by function f_{mc} provides the major contribution to the effectiveness.

To further complement our evaluation, we also compare our method against CMOLRS [130], which formulated the margin as an original-feature driven margin that is *fixed during training*, i.e. using only a simplified version of f_{ms} factor of SAM. In table 3.2, we observe that on the Wikipedia dataset, CMOLRS achieved a mAP@100 of 0.413 while SAM achieves a mAP@100 of 0.541. As the authors of CMOLRS only report mAP@100, we did not included it in table 3.1.

SAM formulates the adaptive margin function f_m, as a dynamic function which gradually enforces triplet-specific margin constraints during training. This confirms the importance of dynamically adjusting margin values during training and of the novel cluster formation and preservation component f_{mc}.

3.6.1.4 Large-scale NUS-WIDE dataset

To further explore the generalization of SAM algorithm, we evaluated SAM in the large-scale full NUS-WIDE dataset, under the same conditions of NUS-WIDE-10k: $\lambda = 0.05$, $f_a = 0.9$ and $k = 0.1$. With larger datasets, more triplet constraints are enforced per epoch. Namely, given a dataset of size N and mini-batches of size b, a total of $N \times b$ triplets are enforced. Thus, by definition, the number of constraints that are enforced during network training, scales linearly with the dataset dimension.

Table 3.3 supports the same conclusions that where drawn from the previous analysis. Namely, SAM outperformed all the compared baselines. As observed previously, starting

Table 3.3: mAP results on the NUS-WIDE dataset.

	NUS-WIDE		
Methods	$I \mapsto T$	$T \mapsto I$	Avg.
CCA [49]	0.244	0.275	0.260
CFA [69]	0.358	0.361	0.360
KCCA [42]	0.348	0.481	0.415
LGCFL [54]	0.512	0.600	0.556
JRL [141]	0.615	0.592	0.604
Corr-AE [28]	0.391	0.429	0.410
DCCA [133]	0.475	0.500	0.488
CMDN [93]	0.643	0.626	0.635
CCL [92]	0.671	0.676	0.674
SAM	**0.701**	**0.707**	**0.704**

activating the adaptive margin near the last epochs of training yields effective embeddings. Moreover, giving more importance to the cluster enforcement and preservation component yields better embedding structure, as also observed in all the three previous datasets. It is also noticeable, that all models improved thanks to the larger training dataset.

3.6.1.5 Overview.

In overall, our method has proven to be effective, outperforming previous state-of-the-art methods on all datasets. The cluster enforcement and preservation component (f_{mc}) proved to be crucial to achieve state-of-the-art performance. Unlike most methods, which impose extra constraints by augmenting a projection network by adding additional loss terms, our approach imposes those constraints by directly adapting the margin between instance pairs during training, thus resulting in a simpler but effective model.

By modeling the semantic inter-category pairwise correlations, our model is able to transfer semantic correlations from the original feature space directly to the common embedding space. Then, by enforcing cluster formation after achieving a stable embedding space organization, our method improves significantly the state-of-the-art.

3.6.2 Scheduled Adaptive Margins Analysis

In this section we examine the behavior of the scheduled adaptive margins, on the NUS-WIDE-10k dataset, (using $f_a = 0.4$, $\lambda = 0.05$ and $k = 0.1$), by looking at the margin values imposed by the model on each triplet constraints, over each epoch (t).

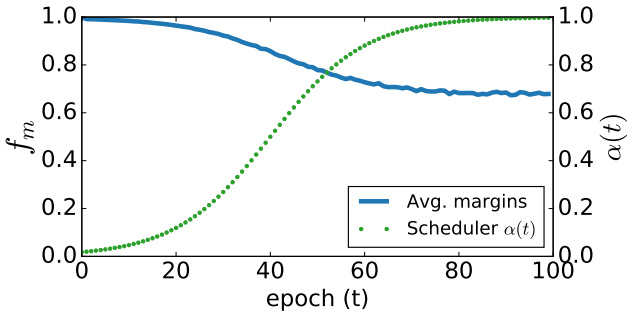


Figure 3.4: Global average adaptive margin f_m over training epochs (t) on the NUS-WIDE-10k. The left y-axis corresponds to the f_m value and the right y-axis to the scheduling function $\alpha(t)$ value.

3.6.2.1 Average margin vs. scheduler function

The scheduler function $\alpha(t)$ shifts from a high-magnitude constant margin ($m = 1$), to the adaptive margin f_{am}. To inspect this behavior, we computed the average margin value, imposed to all triplets, on each epoch t, on the NUS-WIDE-10k dataset. Figure 3.4 shows the average f_m value (blue line) versus the scheduler function value $\alpha(t)$ (green line), over the training epochs. It can be observed that at each epoch, the average margin imposed by f_m tends to be smaller. One can also observe that $\alpha(t)$ has a sigmoidal shape.

3.6.2.2 Average margin values for each Category

In order to provide a deeper understanding of what the model achieves, we show in Figure 3.6, also on the NUS-WIDE-10k dataset, the average margin values between three pairs of categories at each training *epoch t*, and a projection of the final cross-modal embedding space.

The scale of the average margin values in the last epoch ($t = 100$), between each pair of the considered categories, is reflected in the obtained embedding space. It is noteworthy to say that the magnitude of the value m reflects the difference between similarities of pairs of instances, not distance on the embedding space. Nevertheless, the magnitude of the values still allow to confirm its impact in the embedding space organization. For instance, in the plot of Figure 3.6, it can be seen that in the last epochs, our model enforced an average margin of roughly 0.6 between instances of category *window* versus category *sky*, which is much smaller than the value between instances of *window* and *grass*, which is roughly 0.77. Looking at the t-SNE [77] projections, we can actually see that the organization of instances respects these values, with vectors of instances of category *window* having similar directions, in comparison to instances of *sky* than to *grass*.

These experiments are crucial to understand the underpinnings of SAM: Figure 3.5

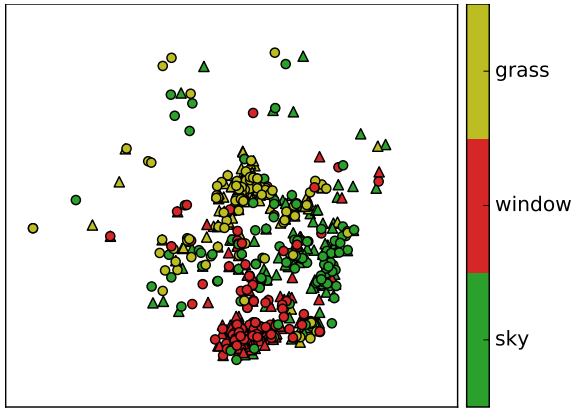


Figure 3.5: t-SNE projections - Scheduled Adaptive Margins between 3 categories.

and Figure 3.6 confirms that the average margin value gradually decreases during training, with triplet constraints over *window-sky* categories having lower magnitude margins than *window-grass*, thus reflecting visual and textual semantic similarity as intended.

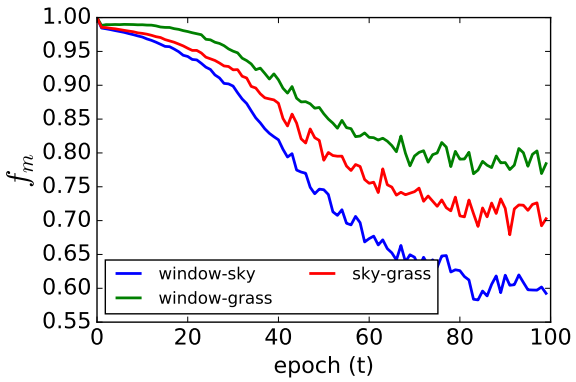


Figure 3.6: Analysis of the margin values over each epoch (t), between three categories.

Figure 3.7 delves into this question and shows the average margin value per category imposed by f_m, against triplets of the remaining categories, at each epoch t. Given the target category c_1 of each plot, each line corresponds to a category c_2. Namely, it corresponds to the average of the margin values, imposed by f_m, to triplets with the positive instance belonging to category c_1 and the negative belonging to category c_2. It is interesting to note that all margins are significantly different. In particular, categories *grass* and *person* are the ones with most homogenous margins. In contrast, categories *sky* and *animal* took full advantage of the scheduled adaptive margins and ended up with very different margins (smaller) to all other categories.

Table 3.4: Analysis of the scheduler and f_{mc} impact.

Method	Pascal Sentences			NUS-WIDE-10k			Wikipedia		
	$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg	$I \mapsto T$	$T \mapsto I$	Avg
SAM ($\alpha(t) = 1$, $\lambda = 1$)	0.586	0.590	0.588	0.539	0.559	0.549	0.406	0.382	0.394
SAM	0.637	0.643	0.640	0.563	0.594	0.579	0.518	0.457	0.487

3.6.2.3 Scheduler and f_{mc} impact

The scheduler, together with the cluster formation and enforcement f_{mc} component of the adaptive margin, are key novel components, responsible for achieving state-of-the-art performance. To confirm this, we evaluated SAM with the scheduler deactivated ($\alpha(t) = 1$) and with f_{mc} disabled ($\lambda = 1$). As can been seen from table 3.4, this results in a drop of performance of $\approx 8\%$, $\approx 5\%$ and $\approx 19\%$, on Pascal Sentences, NUS-WIDE-10k and Wikipedia, respectively, confirming the crucial importance of the scheduler and f_{mc}.

3.6.3 Analysis of Activation Phase f_a and λ Impact

In this section we will analyze the impact of the activation phase f_a and the semantic correlation vs. cluster enforcement trade-off λ parameter. To do this, we measure the *mAP* score on the Pascal Sentences dataset. Namely, we evaluate the activation factor $f_a \in \{0.0., 0.2, 0.4, 0.6, 0.8, 1.0\}$ and trade-off between f_{ms} and f_{mc} (eq 3.8), $\lambda \in \{0.0, 0.1, 0.25, 0.75, 1.0\}$, fixing all the remaining parameters, and show the results in Figure 3.8. The x-axis represents the value of f_a and the y-axis the *mAP* score obtained. Each curve corresponds to a value of λ.

The first observation is that imposing the adaptive margin too early is bad. For instance, when f_a is close to zero, the method starts using the adaptive margin from the beginning of the training, resulting in low performance. This confirms our intuition that in the first few training iterations, the embedding space is still coarsely organized. As the parameter f_a increases, we can see that the results improve significantly, reaching a performance peak on $f_a = 0.4$ (around epoch 40), for four of the five experimented values of λ. Namely, smoothly activating the adaptive margin with $f_a = 0.4$, and giving around 75% weight to f_{mc} (cluster formation and preservation component) and the remaining to f_{ms}, leads to the best performance. For all values of λ, activating the adaptive margin too late leads to significant performance drops. This is due to the fact that by activating later, the network has more chances to overfit using a static margin. At this point, neither the cluster formation f_{mc}, nor the semantic correlations f_{ms} components are able to improve the embedding space organization. Regarding the trade-off parameter λ, we observe the trend that cluster formation has a higher impact on achieving better performance

than semantic correlation, with peak performance occurring when both components are active.

3.6.4 Qualitative Analysis

In this section we qualitatively evaluate SAM. We start by visualizing the obtained embedding space in section 3.6.4.1, and then by performing a success and failure analysis through the inspection of retrieval results, in both the $T \mapsto I$ and $I \mapsto T$ directions, in sections 3.6.4.2 and 3.6.4.3, respectively.

3.6.4.1 Embedding Space Visualization

To complement our quantitative analysis, we perform a qualitative analysis by visualizing the obtained embedding space for our top-performing model, on the NUS-WIDE-10k dataset. As the projection dimension D is set to 200, we apply t-SNE to visualise the obtained data projections on the test set. We randomly sample 500 points per modality. Figure 3.9 shows the resulting space. The figure shows the projection visual (circles) and textual (rectangles) elements over the 10 categories of the dataset, with a different color being associated to each category.

First, we can see that SAM was able to effectively project visual and textual modalities of the same category close to each other. Additionally, one can observe well-defined category clusters. A closer look to the resulting embedding space organization reveals very interesting insights. According to our intuition, semantically correlated categories are actually placed close (similar directions) to each other. For instance, projections of elements from the category *clouds* are very close (in fact mixed) to elements of the category *sky* and *window*. In fact, from a visual and textual perspective, there are images and texts that will actually belong to all of the 3 categories. This is a consequence of the component f_{ms} of the semantic margin formulation. A more fine-grained observation also allows us to observe that some of the mistakes of the model are due to the existence of visual elements with overlapping categories. For instance, for the category *animal*, we can observe that one image was badly projected towards the centroids of *sky* and *clouds* categories. An example of such an image would be a *bird flying*.

3.6.4.2 Success and Failure Analysis - Text to Image

In Figure 3.10, we show the results of SAM, on a set of sampled queries, for the $T \mapsto I$ direction. Images that are relevant have a green border, and images that are non-relevant have a red border.

For example, in query 7 (second row) comprised by a text belonging to the *animal* semantic category, SAM retrieves only images depicting from the same category. The same happens for query 2, which focus on content from category *food*. Interesting, in query 1 (third row), the second image is wrong. While the image depicts a butterfly on a flower, it is annotated as belonging to the category *animal*. However, the text from that query belongs to the category *flower*. This evidences that SAM can capture the semantics of images with multiple concepts, and the retrieved image could in fact be correct.

3.6.4.3 Success and Failure Analysis - Image to Text

In Figure 3.11, we show the results of cross-modal retrieval, on a set of sampled queries, for the $I \mapsto T$ direction. Texts that are relevant have a green border, and texts that are non-relevant have a red border.

As in the results from the previous section 3.6.4.2, we observe the same pattern. For instance, in query 21 (third row), comprised by an image of a building, both the first, the third and the fifth texts contain words that are associated with the image: windows, buildings, architecture. In fact, the second text, which is correct, contains these words.

This leads us to conclude that most SAM mistakes are due to the fact that the NUS-WIDE-10k dataset is multi-class (single category for each image/text), and this is not fine-grain enough for the type of images and text that the dataset comprises. Namely, SAM performs quite well at structuring instances based on their semantics. Even though some retrieved instances are marked as non-relevant, they could in fact be deemed as relevant, depending on the user intent.

3.7 Critical Summary

In this chapter we described a novel method to learn cross-modal embeddings. The method introduces a scheduled activation of adaptive margins that allow for triplet specific margins. The key takeaways of the proposed method are:

- **Adaptive margin constraints**: our approach impose general constraints while training the model by adapting the margins between instance pairs. This overcomes the fact that using a unique margin for all pairs is insufficient to adequately structure the embedding space.

- **Effective learning of pair-specific margins**: results show that adaptive margins deliver state-of-the-art results. This is further possible due to the pair-specific margins that are learned by the model as illustrated by experimental results.

- **Scheduled learning**: new neural-network training approach was introduced that progressively activates the adaptive margin function, through an epoch-aware scheduling strategy.

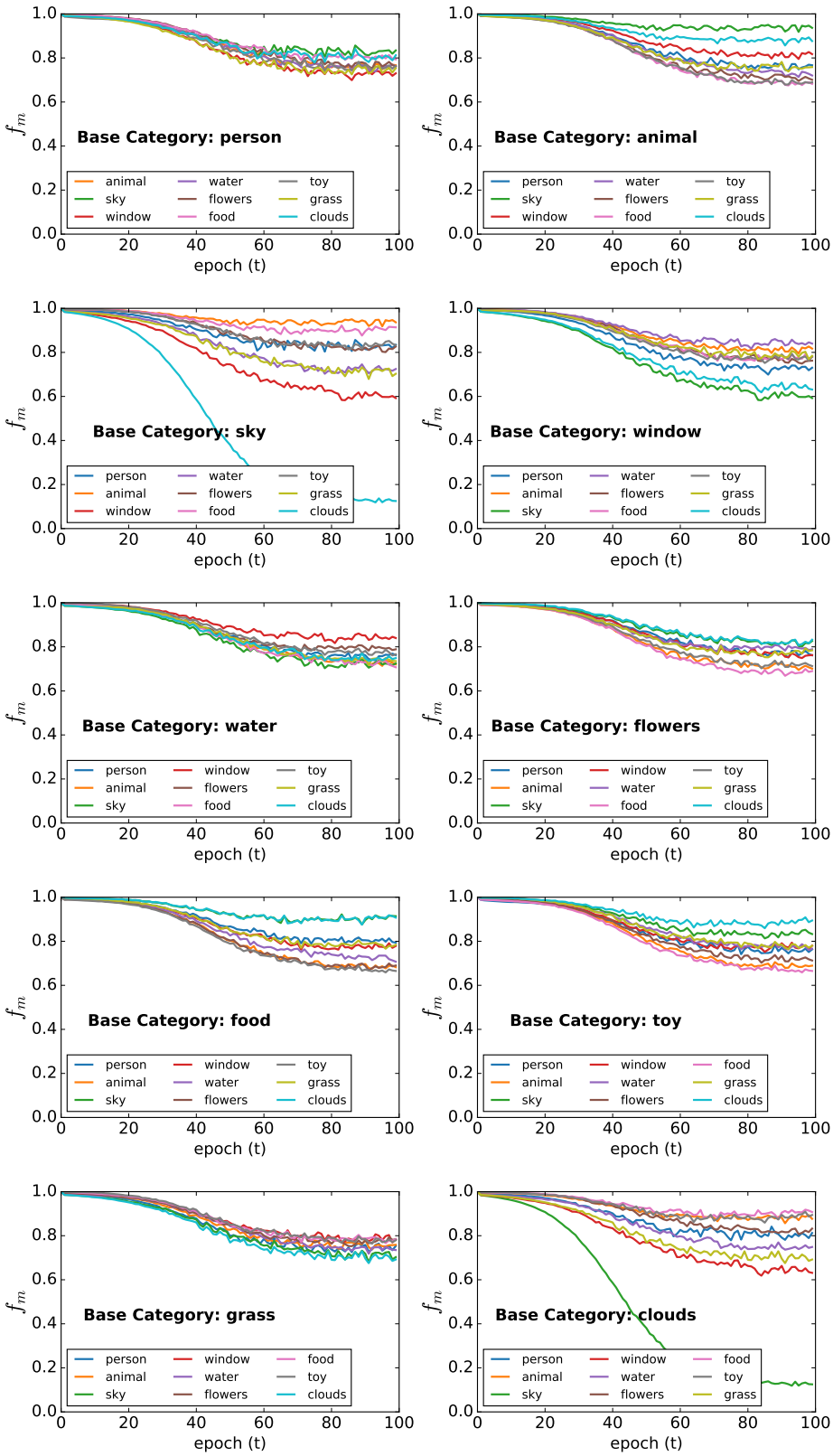


Figure 3.7: Average per-category margin for each category, at each training epoch (t). Average value of f_m between every instance d^i, against all instances d^n of other categories, on NUS-WIDE-10k.

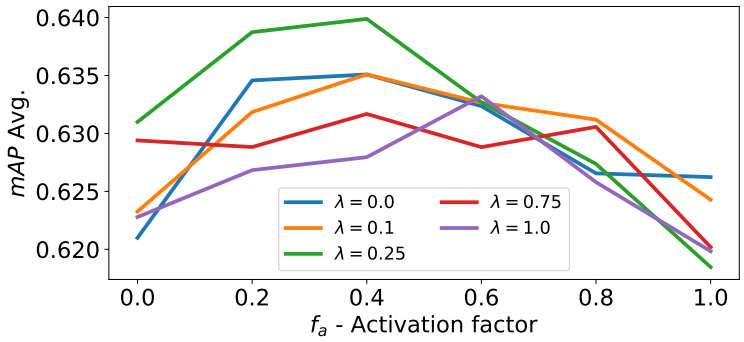


Figure 3.8: Parameter Analysis (λ and activation function f_a) on Pascal Sentences dataset.

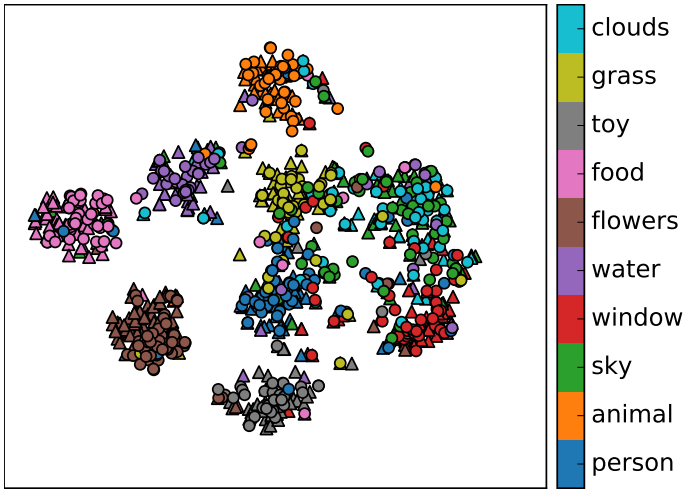


Figure 3.9: t-SNE Visualization of test instances projections of the NUS-WIDE-10k dataset, on the obtained embedding space. Triangles and circles refer to image and text elements, respectively. Best viewed in color.

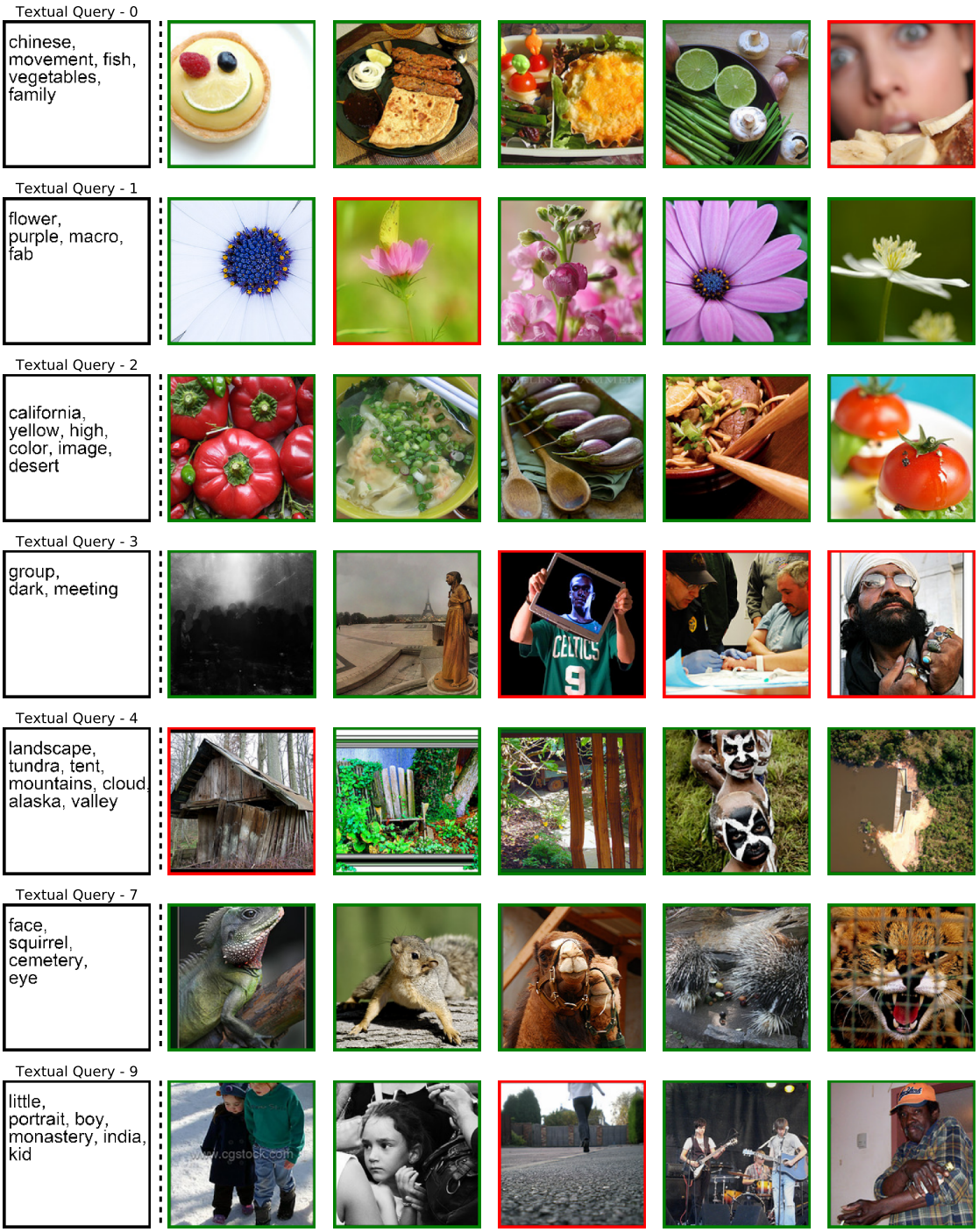


Figure 3.10: Results for query X in the $T \mapsto I$ task. Green border for correct and red for incorrect.

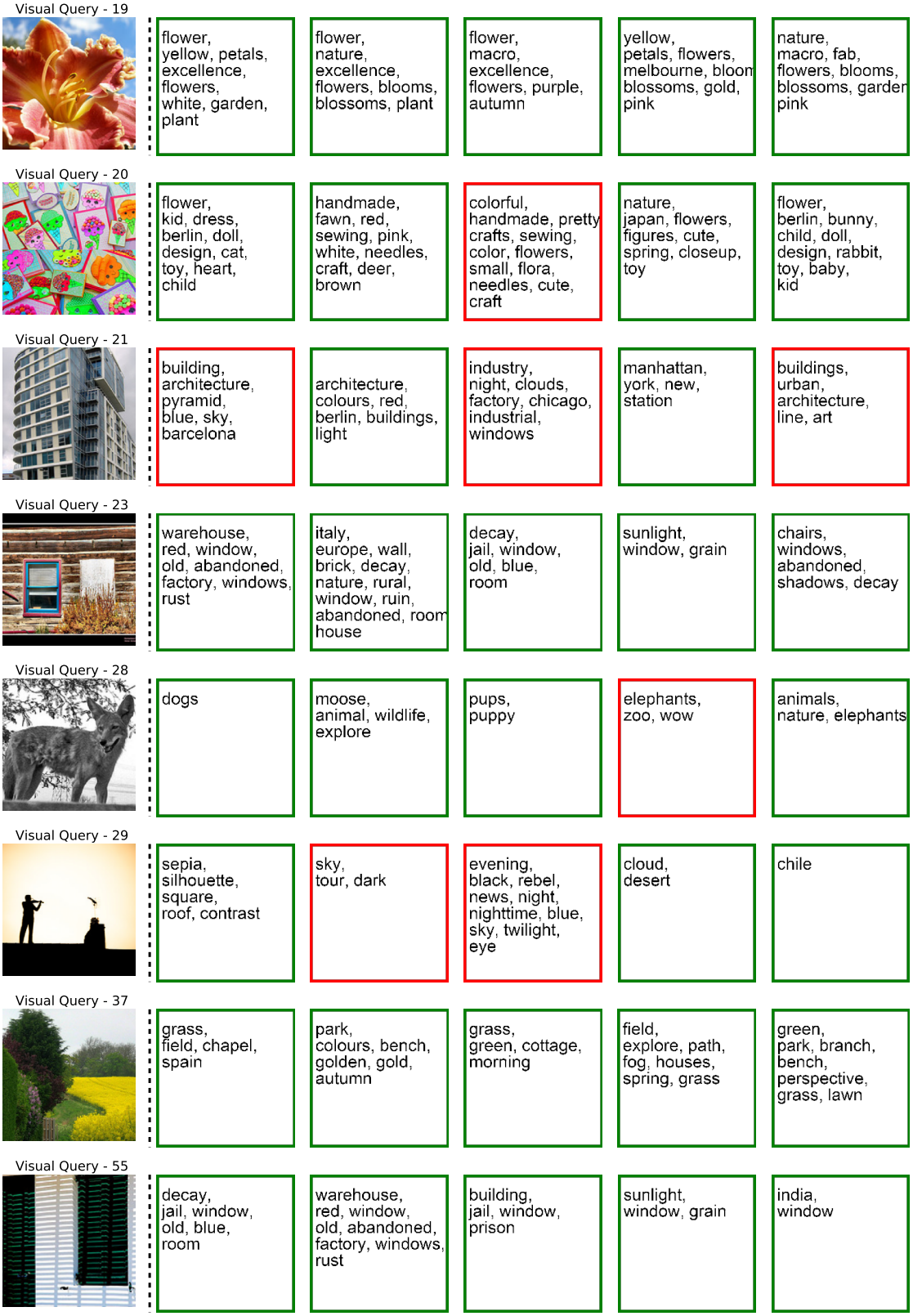


Figure 3.11: Results for query X in the $I \mapsto T$ task. Green border for correct and red for incorrect.

CHAPTER 4

Temporal Cross-modal Embeddings

When learning cross-modal embeddings, the goal is to learn a common space, for visual and textual information, in which its structure reflects how the two modalities are correlated. In some domains, visual and textual patterns of interactions are subject to change over time, which implies the existence of different distributions underlying data: spike-based (single mode), recurrent (multiple modes), etc. One such example is web content, that as discussed in section 1.1, is a mirror of real life: follows emerging topics and trends, with visual content and their descriptions reflecting how people interpreted and reacted to a given topic.

As discussed in section 2.2.5.4 of Related Work, a **common assumption of cross-modal embedding learning works is that corpora is static**. Consequently, temporal correlations between *visual-textual* pairs have been overlooked.

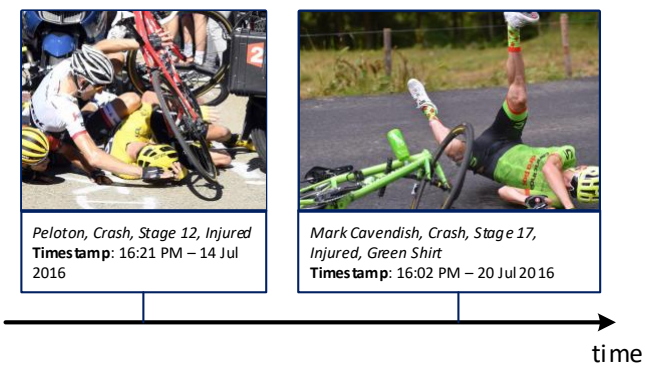


Figure 4.1: Temporal dynamics of content from the semantic category *Crash* (Tour-de-France 2016), and temporal pairwise variations with corresponding visual elements.

Looking at Figure 4.1 one can see two visually similar images, *i.e.* both depict a cyclist falling. Namely, they have the same semantic category (*crash*), but occurred at two distinct instants in time, thus refer to different crashes. We know this since the textual descriptions, refer to different cyclists and places, indicating a **semantic context change**. Specifically, the correspondences between the visual materialization of the concept *crash*, and text, in the domain of the *Tour de France* topic, changed. These context changes lead to the existence of cross-modal pairwise correlations that change over time, *i.e.* correlations relative to each pair of image-text. As static cross-modal embedding models neglect temporal information, and focus solely in structuring instances based on their semantic category, the two images from figure 4.1 would be structured in the same neighborhood (maybe even in the same point on the manifold).

As discussed in section 2.3, numerous works [10, 58, 66, 75, 82, 103, 117, 119] have leveraged on the dynamics of web content for diverse tasks. Namely, they exploit the fact that content from dynamic collections, from certain topics, follows some temporal pattern. The take home message is that data temporal insights proved to be crucial to increase the discriminative power. Therefore, it follows that the temporal dimension should be accounted, such that:

a) Latent *visual-textual* temporal correlations along the collection time span are captured and quantified;

b) Data is structured in the cross-modal embedding space according to semantic and temporal correlations.

In this chapter, we introduce a model that departs from previous static cross-modal embedding learning works, by devising a temporal cross-modal embedding learning model, that accounts for the aforementioned aspects.

4.1 Formulating the Temporal Embedding Space Hypothesis

The main hypothesis we exploit in this chapter is that pairwise visual-textual patterns of interaction change over time. This is supported by the existence of dynamic visual-textual pairs (Figure 4.1), originating changes in cross-modal correlations among the visual and textual dimensions of the problem's data.

When learning an embedding space that bridges vision and language in a supervised setting (instances are labeled), we are interested in retaining similarity between visual and textual elements that are semantically correlated. Thus, if any two elements (image

or text) are not semantically related, then temporal correlation should not be accounted. Consequently, we argue that temporal correlations between instances of a same semantic category should lead to the investigation of new embedding spaces that capture such data interactions.

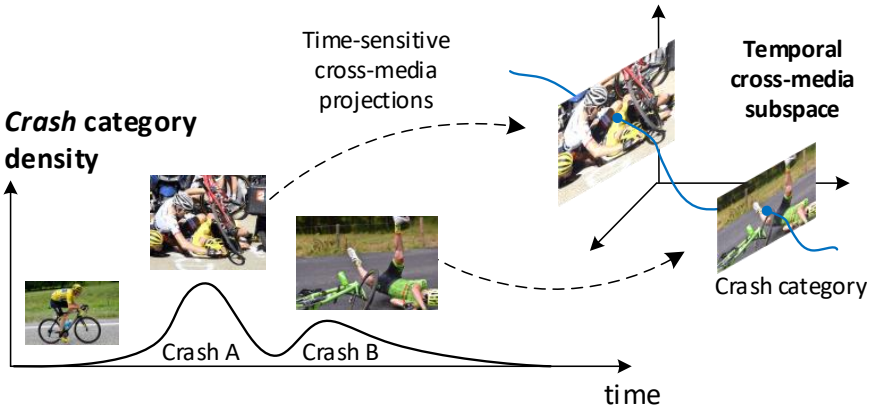


Figure 4.2: Temporal dynamics of semantic category *Crash* (TDF2016), and temporal pairwise variations with corresponding visual elements.

To illustrate this, consider the illustration of Figure 4.2. The plot depicts the temporal density (number of instances per instant) of all content of the crash category, from Tour de France 2016. We highlight the existence of two modes, each corresponding to the two crashes that took place at different moments in time.

Given that content from *Crash A* and *Crash B* belong to the same semantic category *crash*, static cross-modal embeddings would ignore the fact that two distinct crashes happened, and consider instances from both crashes as semantically identical. This means that in the optimal structuring, content from both crashes would be structured in the same neighborhood. Formally, the model would structure content from both crashes such that the similarity between any two instances x_*^1 and x_*^2 is maximal ($s(x_*^1, x_*^2) \rightarrow 1$). As a consequence, any information regarding data original temporal correlations is lost. Instead, in this chapter we seek for a model that accounts for this information.

4.1.1 Modeling Relative Temporal Correlation

Modeling temporal correlations raises many challenges for cross-modal embedding learning methods. The **relative temporal correlation**, between two instances, may be governed by different distributions on different collections. Figure 4.3 extends the previous illustration from Figure 4.2, to show the possibility of using different models (with each possibly following different distributions) for estimating relative temporal correlation.

The first challenge that arises is the modeling and quantification of temporal correlations. Given the temporal distribution of content from the semantic category *crash*, we

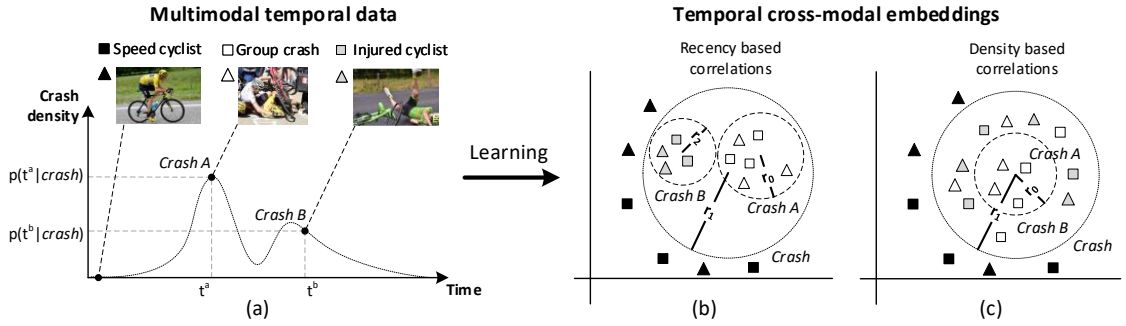


Figure 4.3: Temporal correlations of same-category multimodal data (on the left) follow an unknown density distribution. The temporal cross-modal embedding (on the right) captures these temporal correlations by organizing projected data accordingly, for each specific semantic category.

show at the right, in Figure 4.3, the desired final embedding structures, when assuming either one of the two types of temporal distribution:

Recency-Based - Figure 4.3 (b) - Temporal correlation stems from temporal proximity. Thus, instances are temporally correlated if they are close in time;

Density-Based - Figure 4.3 (c) - Temporal correlations stem from estimated temporal density[1]. Instances are temporally correlated if their density, on the time instant corresponding to their timestamp, is similar.

To further illustrate these two types of temporal correlation, consider the following. Let r_1 define the neighborhood size of instances from the semantic category *crash* in the embedding space. In recency-based correlations, Figure 4.3 (b), content from each crash is structured such that *Crash A* and *Crash B* fall within none intersecting neighborhoods, of size r_0 and r_1, respectively. In density-based correlations, Figure 4.3 (c), content from *Crash A*, on the top of the corresponding highest crash peak, would be structured in the same neighborhood r_0. The most suited temporal correlation should depend on the collection temporal patterns. The choice of a type of temporal correlations implies a commitment in terms of expressiveness of the model in capturing temporal correlations. While recency-based correlations are suited to structure data that occurred only once, and at specific moments in time, it fails to capture data with multiple modes. In such scenario, density-based correlations should be used instead.

We tackle the challenges presented in this section by defining a cross-modal embedding space where semantically similar and *mutually temporally related instances*, lie in the same neighborhood. The goal is to learn effective projections where cross-modal

[1]Temporal density may be estimated from any temporal signal (*e.g.* documents frequency over time instants, etc.).

patterns are captured and perturbed according to pairwise (relative) temporal correlations. Different distributions, underlying relative temporal correlation, will be studied and used to structure multimodal data.

4.2 Embedding Definition

This section will detail the proposed Temporal Cross-modal Embedding, which we refer to as TempXNet, *i.e.* its definition and neural architecture. The goal is to design a temporal cross-modal neural architecture, to learn projections for both textual (f_T) and visual (f_V) data, while modeling temporal correlations between modalities. A temporal embedding learning approach will be devised to obtain *Time-sensitive modality projections*, through the enforcement of temporal constraints between semantically similar instances. This novel embedding space enables effective retrieval in a temporally-aware cross-modal embedding.

In this chapter, we consider the two types of temporal correlations presented in section 4.1, to model the underlying dynamics of instances: **Recency-based** and **Density-based**. These will model intra-category temporal correlations at two levels of granularity: at the documents' timestamp level and at individual words' level. The key aspect of the TempXNet is that the proposed model will be flexible enough to support cross-modal temporal correlations following parametric, non-parametric and latent-variable distributions. The temporal cross-modal embedding will now be formally defined in the next section.

4.2.1 Temporal Cross-modal Space

We start by recapping the notation introduced in section 1.1.3 and defining the task of temporal cross-modal embedding learning. Let $C = \{d_i\}_{i=1}^{N}$ be a set of N *visual-textual* instance tuples

$$d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, ts^i, c^i). \tag{4.1}$$

where $\mathbf{x}_V^i \in \mathbb{R}^{D_V}$ and $\mathbf{x}_T^i \in \mathbb{R}^{D_T}$ are the feature representations of the image and textual elements, respectively. The collection timespan is defined by $TS = [t_{start}, t_{end}]$, where t_{start} and t_{end} are the first and last instants of the dataset, respectively. All instances are timestamped, with $ts^i \in TS$ denoting the timestamp of an instance d^i. In this chapter, c^i denotes the set of semantic categories. It follows that each instance d^i can be associated with one or more categories. A Temporal Cross-modal embedding space is formally defined as follows:

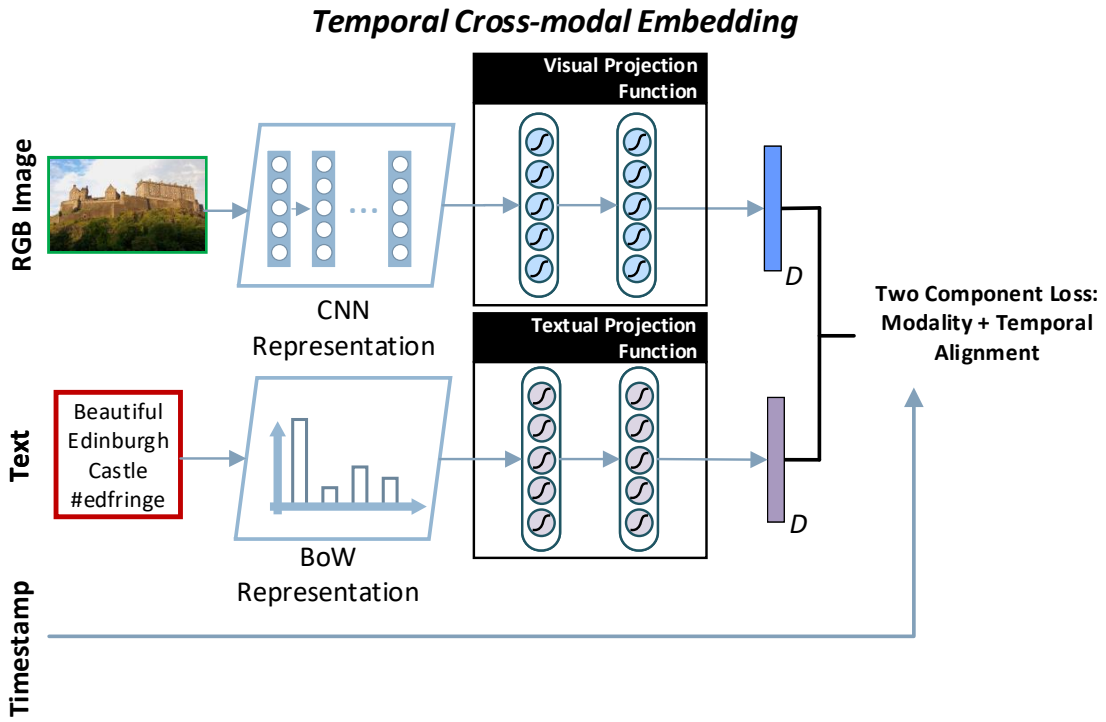


Figure 4.4: Temporal cross-modal embedding learning overview. Visual (blue) and textual (purple) instances are mapped to a D dimensional cross-modal space.

Definition of Temporal Cross-modal Embedding

Definition 1. A *Temporal Cross-modal Embedding space* refers to a common embedding space $\mathcal{S} \in \mathbb{R}^D$, learned from a timestamped collection C, that structures visual and textual elements of data instances according to their semantic category c^i and pairwise temporal correlations, measured by a function $f_t(\cdot)$, across different modalities.

4.2.2 Time-sensitive Cross-modal Neural Projections

Given the aforementioned definition, it follows that both $\mathbf{x}_V$ and $\mathbf{x}_T$ original spaces are dissimilar and obtained without accounting for time. Namely, they denote heterogeneous information sources, as each space may have different dimensionality, semantics and distributions, making them incompatible. This leads us to the projections,

$$f_V(\cdot;\boldsymbol{\theta}_V) : \mathbb{R}^{D_V} \mapsto \mathbb{R}^D \qquad \textit{and} \qquad f_T(\cdot;\boldsymbol{\theta}_T) : \mathbb{R}^{D_T} \mapsto \mathbb{R}^D \tag{4.2}$$

mapping images $\mathbf{x}_V^i$ and texts $\mathbf{x}_T^i$ to a common temporal cross-modal embedding, with dimensionality D, according to the image and text projection functions f_V and f_T, with parameters $\boldsymbol{\theta_V}$ and $\boldsymbol{\theta_T}$, respectively.

Both projection functions f_V and f_T do not take time as input. Instead, as will be described in section 4.2.3, information regarding temporal correlations will be used by the model loss function to structure embeddings accordingly. Projections are time-sensitive as images and texts are projected to regions of the embedding space where semantically similar and temporally correlated elements lye together (*e.g.* the two images from Figure 4.1 should lye in different regions under a recency-based model). In practice, after training, the time dimension is discarded. The implications of this aspect will be discussed throughout the remainder of the chapter. The resulting embeddings, produced by f_V and f_T will be characterized by the properties defined in the following section.

4.2.3 Embedding Properties

In this section we state the fundamental properties that define the structure of a temporal cross-modal space, and that projections f_V and f_T need to satisfy. These aim to materialize an embedding space that complies with definition 1. The properties are:

- **Property 1**. Two elements will be maximally correlated (high similarity) in the new embedding space, *i.e.* projected to the same neighborhood, if they share at least one semantic category and if they are strongly correlated in time;
- **Property 2**. Considering the same semantic category, temporally correlated instances will lie in the same fine-grain neighborhood, while temporally uncorrelated instances are expected to lie in different neighborhoods (*i.e.* lie far apart);
- **Property 3**. Two elements will be minimally correlated (low similarity) in the new embedding space if they do not share any semantic category;

Property 1 and **Property 2** define the intra-category structure of the embedding, *i.e.* how instances of the same category should be organized. These are demonstrated in Figure 4.3, where images and texts from the same semantic category, are structured according to a given type of temporal correlation: Recency-based **(b)** or Density-based **(c)**. Namely, intra-category embedding space organization will be perturbed by temporal correlations that are captured by a function f_t. The function f_t will be grounded on a temporal distribution θ_{temp}, which can follow a parametric, non-parametric or latent-variable model. Finally, **Property 3** defines the inter-category structure of the embedding, by stating that instances from different semantic categories should be far apart, regardless of their temporal correlations.

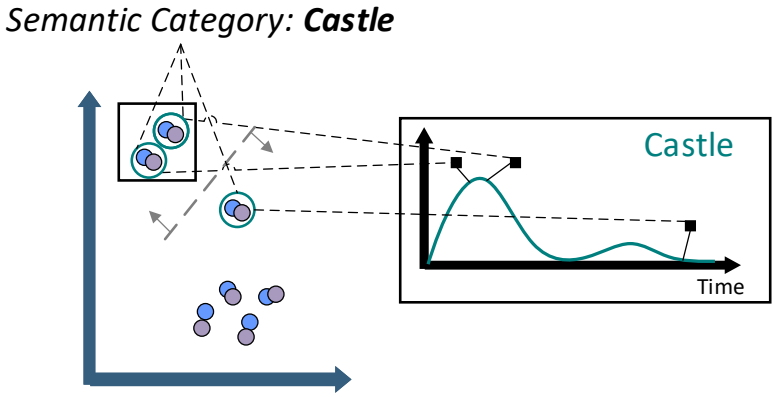


Figure 4.5: In terms of intra-category structuring, the space is perturbed to approximate temporally correlated instances, and to separate uncorrelated ones.

The rationale enforced by **Property 1** and **Property 2** is depicted in Figure 4.5. In the figure, within the content of the semantic category *castle*, estimated temporal correlation, shown in the plot at the right, is used to structure data in the embedding space.

4.3 Temporal Embedding Model Design and Learning

The goal of temporal cross-modal embedding learning is to create a new embedding where semantic and temporal latent correlations, for instances of the same category, are represented at essentially two granularity levels: a) inter-modality and intra-category pairwise correlation (Properties 1 and 2), and b) inter-category correlation (Property 3).

On the obtained temporal embedding, the encoding of the temporal dimension is achieved by smoothing the visually-textually aligned embedding space with a set of temporal constraints imposed on the model loss function.

4.3.1 Joint Temporal Triplet Ranking Loss

To learn the time-sensitive cross-modal projections $f_V(\cdot;\boldsymbol{\theta}_V)$ and $f_T(\cdot;\boldsymbol{\theta}_T)$, it is essential to maximize correlation in the new embedding space between the two modalities, both at the semantic and temporal dimensions. Thus, the projections into the temporal cross-modal embedding need to capture the temporal traits of semantic categories, which are grounded on temporal correlations across visual and textual modalities. In practice, we argue for projections that are learned with an objective function of the form

$$\arg\min_{\boldsymbol{\theta}_V,\boldsymbol{\theta}_T} \mathcal{L}(\boldsymbol{\theta}_V,\boldsymbol{\theta}_T) \tag{4.3}$$

where $\mathscr{L}$ corresponds to a cross-modal loss that maximizes the similarity over semantically similar representations and minimizes the similarity between semantically dissimilar instances' representations (Property 3).

It is crucial to learn effective projections, that map original modality vectors to a new space where pairwise (visual and textual modalities) and instance's semantic correlations are represented. As discussed in section 2.2.5 of the related work chapter, the triplet loss is among the top performing loss functions for cross-modal representation learning [121, 124, 135]. Therefore, we formulate $\mathscr{L}$ using the triplet-loss,

$$\begin{aligned}\mathscr{L}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = &\sum_{i,n} max(0, m - s(\mathbf{x}_V^i, \mathbf{x}_T^i) + s(\mathbf{x}_V^i, \mathbf{x}_T^n)) \; + \\ &\sum_{i,n} max(0, m - s(\mathbf{x}_T^i, \mathbf{x}_V^i) + s(\mathbf{x}_T^i, \mathbf{x}_V^n)),\end{aligned} \tag{4.4}$$

where $\mathbf{x}_V^n$ and $\mathbf{x}_T^n$ are images and texts representations from negative instances, w.r.t. an instance d^i. To effectively capture inter-modality correlations, we enforce a set of triplet constraints on both modality directions: Image to Text and Text to Image, corresponding to the first and second terms of eq. 4.4. As detailed in section 1.1.3, similarity between projections is computed by a dot product over two unit-norm, ℓ_2 normalized vectors, making it equivalent to cosine similarity.

Then, we subject the cross-modal loss to temporal smoothing constraints, imposed by a temporal factor $\mathscr{L}_{temp}$, grounded on a temporal model $\boldsymbol{\theta}_{temp}$. Due to the stochastic nature of neural networks, it is hard to define an optimization objective (or even infeasible) in which we can incorporate a set of constraints and make sure that the number of violations to these constraints will be zero. This would require different projection functions, with the constrained formulation of the objective function $\mathscr{L}_{temp}$ solved to optimality, *i.e.* solve the problem

$$\arg\min_{\boldsymbol{\theta}_V, \boldsymbol{\theta}_T} \mathscr{L}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) \;\; \text{s.t.} \;\; \mathscr{L}_{temp}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = 0. \tag{4.5}$$

Instead, we devise a softly-constrained objective. Namely, temporal constraints $\mathscr{L}_{temp}$ are relaxed as an additive smoothing term, added to $\mathscr{L}$:

$$\arg\min_{\boldsymbol{\theta}_V, \boldsymbol{\theta}_T} \mathscr{L}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) + \lambda \cdot \mathscr{L}_{temp}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T, \boldsymbol{\theta_{temp}}), \tag{4.6}$$

where λ is an hyper-parameter that may be optionally used to control the influence of the temporal factor.

The temporal factor term $\mathscr{L}_{temp}$, detailed in section 4.3.2, is backed up by a temporal model θ_{temp}, estimated from the corpus, that covers two temporal aspects:

1. Instances' *temporal signature* over the corpus C time span;

2. Smoothed temporal correlation functions, based on the aforementioned temporal signatures.

The objective function from eq. 4.3 leads to cross-modal projections fundamentally different from previous works, as in these works, images and text are grouped in a temporally agnostic manner.

4.3.2 Temporal Cross-modal Soft-Constraints

Temporal embedding learning properties are enforced over semantically similar instances only, through a set of soft-constraints. Thus, the temporal factor $\mathscr{L}_{temp}$ is defined as:

$$\mathscr{L}_{temp}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T) = \sum_i \mathscr{L}_{temp}(d^i; \boldsymbol{\theta}_V, \boldsymbol{\theta}_T), \tag{4.7}$$

From equation 4.6, $\mathscr{L}_{temp}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T)$ is added to eq. 4.3 as a smoothing term. The rationale of equation 4.7 is to smooth the model by constraining the learned projections for every instance d^i, with temporal soft-constraints. We stress that $\mathscr{L}_{temp}(\boldsymbol{\theta}_V, \boldsymbol{\theta}_T)$ is used to perform *intra-category structuring*, *i.e.* structure instances of the same category .

As such, for each instance d^i, we formulate two soft-constraints, C_a and C_b, which are combined as:

$$\mathscr{L}_{temp}(d^i; \boldsymbol{\theta_V}, \boldsymbol{\theta_T}, \boldsymbol{\theta_{temp}}) = C_a(d^i) + C_b(d^i). \tag{4.8}$$

Essentially, for a given instance d^i, $\mathscr{L}_{temp}$ will iterate through all the positive instances d^j (sharing at least one semantic category with d^i), and enforce the properties described in section 4.2.3.

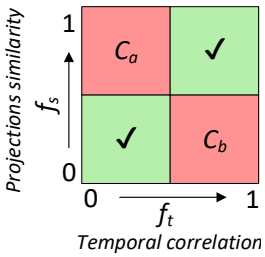


Figure 4.6: Constraints violations rationale.

For a single instance d^i, let $J = \{j : c^j \cap c^i \neq \emptyset\}$ be the set of positive examples d^i of category of c^i. Considering the diagram above explaining the constraints rationale, these are formulated as:

- Temporally correlated instances d^i and d^j, from the same category $c^i = c^j$, with distant cross-modality projections, should have similar projections. Violations to this constraint are captured as follows:

$$C_a(d^i) = \frac{1}{|J|} \sum_{j \in J} \underbrace{f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}})}_{\text{Temporal Correlation}} \cdot \underbrace{(1 - f_s(d^i, d^j; \boldsymbol{\theta_V}, \boldsymbol{\theta_T}))}_{\text{Cross-modal Similarity}}; \tag{4.9}$$

- Temporally uncorrelated instances d^i and d^j, from the same category $c^i = c^j$, with close cross-modality projections, should lie far apart, thus having distant projections. Violations to this constraint are captured as follows:

$$C_b(d^i) = \frac{1}{|J|} \sum_{j \in J} \underbrace{(1 - f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}}))}_{\text{Temporal Correlation}} \cdot \underbrace{f_s(d^i, d^j; \boldsymbol{\theta_V}, \boldsymbol{\theta_T})}_{\text{Cross-modal Similarity}}, \tag{4.10}$$

where $f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}})$, detailed in section 4.4, is a temporal correlation assessment function that evaluates how correlated in time two instances d^i and d^j are. Finally, $f_s(d^i, d^j; \boldsymbol{\theta_V}, \boldsymbol{\theta_T})$, detailed in section 4.3.3, is a cross-modality similarity function that evaluates how close each modality projection is, w.r.t. to the other modality, on the cross-modal embedding. For each instance, we average pairwise violations w.r.t. to each corresponding positive instance, to deal with unbalanced positive sets. From eq. 4.8 the two constraints (C_a and C_b) are assessed by computing the two products between temporal and cross-modality distances.

4.3.3 Cross-modality similarity

Cross-modality similarity f_s, computed over semantically similar instances of d^i, is defined based on the harmonic mean between the cross-modality projections' similarities:

$$f_s(d^i, d^j) = 2 \cdot \left(\frac{1}{f_V(\mathbf{x}_V^i) \cdot f_T(\mathbf{x}_T^j) + \epsilon} + \frac{1}{f_T(\mathbf{x}_T^i) \cdot f_V(x_V^j) + \epsilon} \right)^{-1} \tag{4.11}$$

where again, similarity is computed by a dot product between ℓ_2 normalized vectors. A small constant ϵ is added to the denominator to avoid zero division. Essentially, f_s assesses the alignment between the representations obtained by projections $f_V(\cdot)$ and $f_T(\cdot)$, over two instances, by equally weighting both modalities' projections.

4.4 Temporal Soft-Smoothing Correlation Functions

For semantic categories and words, temporal correlation strength within different instances, is expected to vary. This variation is reflected on the dynamic behavior of content. On a corpus C, such behavior is accounted by $\mathcal{L}_{temp}$, through a temporal correlation assessment function f_t. We materialize the later at two fundamentally different levels: *category* and *word* temporal behavior.

4.4.1 Recency-based Correlations

The rationale of recency-based correlation is to favor instances which are temporally correlated according to temporal proximity. Thus, an instance d^i is temporally correlated to another instance d^j if both occur close in time. For temporally distant ones, a non-linear decay is applied. Given two instances d^i and d^j, and their associated timestamps ts^i and ts^j, respectively, we formulate the Recency-based correlation as:

$$f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}}) = e^{\frac{-|ts^i - ts^j|}{h}} \tag{4.12}$$

where $\boldsymbol{\theta_{temp}} = \{h\}$, with h being a parameter that allows controlling the decay level, according to the granularity and time span of different corpora, and $f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}})$ maps to the range $]0, 1]$.

4.4.2 Category-based Correlations

It is expected that different semantic categories, will have different dynamics w.r.t. to the documents distribution over time. Within the content of a category, there may be different temporal behaviours. Namely, there can be situations in which we want instances that are far apart, but happened at an important/relevant moment[2], to be structured together in the cross-modal embedding space. This is the example of Figure 4.3 which depicts two crashes in TDF2016 and the documents' distribution has two peaks. To structure together content that fall within these two crashes, we can look at the density distribution of documents over time, for a given category.

As such, we propose to assess temporal correlations by directly comparing temporal density distribution ϕ_c of categories. Given ϕ_c, we define the temporal density of $c \in c_i$, at time t, as a probability function $p(t|\phi_c)$. While the probability function $p(t|\phi_c)$ can be estimated in a multitude of ways, we opted for estimating this probability in a straightforward way, which is based on the number of multimodal documents per day.

[2]Importance/relevance stem from the scenario which we are trying to model. In this situation we relate importance to peaks of social media reactions to a topic (user posts).

Accordingly, category-based correlations are then defined as:

$$f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}}) = p(ts^i|\phi_c) \cdot p(ts^j|\phi_c), \tag{4.13}$$

such that $p(t|\phi_c)$ corresponds to the relevance of label $c \in c_i$, at time t. When two instances share more than one label, we consider the value of the label that maximizes f_t. Kernel Density Estimation (KDE), with a Gaussian Kernel, is used to obtain a smoothed estimation of ϕ_c:

$$p(t|\phi_c) = f_{kde}(t, c) = \frac{1}{n \cdot h} \sum_{i=1}^{n} K(\frac{t - t_i}{h}) \tag{4.14}$$

with h corresponding to the bandwidth hyper-parameter, used to control the smoothness of the estimated density. Therefore, we have $\boldsymbol{\theta_{temp}} = \{h\}$. For each category, a KDE model is estimated by running through the set of timestamps $T_c = \{ts^i : c \in c^i,\ d^i \in C\}$.

4.4.3 Topic-based Correlations

One of limitations of the previous type of temporal correlation, is that temporal densities are estimated on category information, which can be, per see, too broad. Individual word's dynamic behavior provides a more fine-grain insight regarding *visual-textual* temporal pair correlations. Namely, it is expected that some *domain-specific* words will have a rich dynamic behavior, depicting temporal correlations, which should be accounted for. Such correlations are also much more fine-grained, when compared to individual semantic categories. Moreover, within the documents' of a category, there may be some words that can act as good discriminators w.r.t. to temporal correlation. Thus, we defined a third type of temporal correlation that has the same rationale as Category-based correlations, but modify it to be more fine-grain.

We model temporal density distributions ϕ_w of each word $w \in \mathbf{x}_T^i$ of a instance d^i, through a dynamic topic modeling approach [15]. Figure 4.7 illustrates the estimated density distributions for four different words. One can clearly see that during the time span of the event (marked by dashed lines), there are in fact different temporal behaviors. For instance, while the word *opening* and *castle* have high density at the beginning of the event, the density of the word *show* is practically constant across the whole event.

Word Topic-based correlations are defined as:

$$f_t(ts^i, ts^j; \boldsymbol{\theta_{temp}}) = \ p(ts^j|\mathbf{x}_T^i) = \prod_{w \in \mathbf{x}_T^i} p(ts^j|\phi_w), \tag{4.15}$$

such that $p(t|\phi_w)$ corresponds to the density of word w, at time t. Equation 4.15 measures temporal correlation by comparing the temporal density of words in d^i, at timestamp

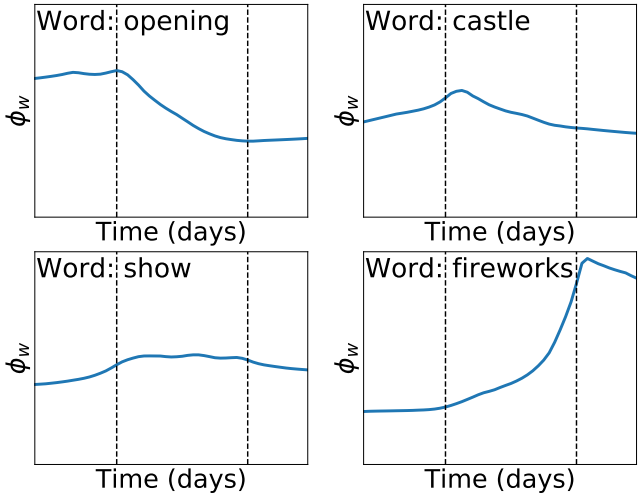


Figure 4.7: Words temporal relevance. Each plot depicts the mean latent-topical temporal curve ϕ_w, over each day, on the Edinburgh Festival dataset. Vertical lines mark the event timespan.

ts^j, that corresponds to the timestamp of document d^j. In other words, eq. 4.15 combines density values of each instance d^i text, through a product. This away, we can assess the relative temporal correlation between two instances d^i and d^j at the word-level.

To estimate ϕ_w, we resort to Dynamic Topic Modeling, specifically D-LDA [15] that intrinsically accounts for the time evolution of latent topics. D-LDA is discussed in section 2.3.2.2 of Chapter 2. It allows the study of the temporal behaviors of individual words. D-LDA is applied to the corpus C with time slices referring to individual days. Then, for each word and latent-topic p, a temporal density curve ϕ_{wp} is estimated. The element-wise mean over all latent-topics is computed as $\phi_w = \sum_{p=0}^{P} \phi_{wp}$, and normalized. Then, for a given word w:

$$p(t|\phi_w) = f_{dlda}(t, w) = \phi_w(t), \tag{4.16}$$

where $\phi_w(t)$ denotes the estimated averaged temporal density, at time instant t, across all topics. Given that we average each ϕ_{wp} over the P latent-topics and that each word w reveals different behaviors on each latent-topic, we obtain a model that captures word variations w.r.t. word correlations with groups of words, over time. Figure 4.7 depicts estimated temporal density curves ϕ_w, revealing the diversity on the different types of dynamic behavior of individual words.

4.5 Neural Model and Architecture

Projections $f_V(\cdot;\boldsymbol{\theta}_V)$ and $f_T(\cdot;\boldsymbol{\theta}_T)$ are each materialized using an independent neural network. Figure 4.4 depicts the neural architecture. Following [25, 88, 133], we consider two neural networks to learn non-linear mappings, with $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_T$, denoting each sub-network's learnable parameters for image and textual modalities, respectively. Formally, the temporal cross-modal projections are defined as:

$$f_V(\mathbf{x}_V^i) = \underbrace{tanh(\boldsymbol{\theta}_{V_2} \cdot tanh\left(\boldsymbol{\theta}_{V_1} \cdot \mathbf{x}_V^i\right))}_{Visual\ Projection}, \qquad f_T(\mathbf{x}_T^i) = \underbrace{tanh(\boldsymbol{\theta}_{T_2} \cdot tanh\left(\boldsymbol{\theta}_{T_1} \cdot \mathbf{x}_V^i\right))}_{Textual\ Projection}, \tag{4.17}$$

in which $\boldsymbol{\theta}_* = \{\boldsymbol{\theta}_{*_1}, \boldsymbol{\theta}_{*_2}\}$,where $\boldsymbol{\theta}_{*_1}$ and $\boldsymbol{\theta}_{*_2}$ correspond to each modality first and second layers weight matrices, respectively. Through the composition of several non-linearities, neural networks are able to model complex latent correlations. Thus, for each modality, a feed-forward network, comprising 2 fully connected layers is used. The first layer has 1024 dimensions and the second one has D dimensions.

Each modality network takes as input the corresponding modality of an instance d^i. Namely, a visual projection sub-network takes as input the image $\mathbf{x}_V^i$ feature representation. For semantically rich image representations, we extract features from a pre-trained convolutional neural network on the task of image classification. Namely, we use a Pre-trained ResNet-50 [45], with the last fully connected layer removed (softmax) to extract features. The textual projection sub-network, takes as input a bag-of-words representation of the text $\mathbf{x}_t s^i$. Both original modality representations are then embedded onto a new D-dimensional embedding space.

Apart from the training phase, both sub-networks are decoupled and thus can be used independently to individually map a single modality.

4.6 Evaluation

In this section we evaluate the temporal cross-modal embedding learning model. We start by describing the dataset in section 4.6.1, the methodology in section 4.6.2, and finally the training and implementation details in section 4.6.3.

Table 4.1: SocialStories dataset information regarding EdFest2016 and TDF2016 events. Seed terms/hashtags, event and crawling time spans are shown.

Event		Keywords	Event Span	Crawling Span
EdFest 2016	Terms	Edinburgh Festival, Edfest, Edinburgh Festival 2016, Edfest 2016	From: 2016-08-04	From: 2016-07-01
	Hashtags	#edfest, #edfringe, #EdinburghFestival, #edinburghfest	Until: 2016-08-24	Until: 2017-01-01
TDF 2016	Terms	le tour de france, le tour de france 2016, tour de france	From: 2016-07-02	From: 2016-06-01
	Hashtags	#TDF2016, #TDF	Until: 2016-07-24	Until: 2017-01-01

4.6.1 Datasets

We consider two datasets: the a) NUS-WIDE benchmark (also used in chapter 3) and b) SocialStories, which was specifically created to comprise content with dynamic behaviour.

4.6.1.1 NUS-WIDE [22]

In chapter 3, we evaluated the proposed cross-modal embedding framework on the task of static cross-modal retrieval. In that chapter, aside from two other datasets, we used as benchmark the NUS-WIDE dataset. Now we are interested in evaluating how the Temporal Cross-modal Embedding structures multimodal documents based not only on semantic but also temporal correlations. NUS-WIDE, which is a standard benchmark dataset used in the cross-media retrieval task, and can be seen as a photo gallery, *i.e.* images are grouped by semantic category, but do not have any explicit temporal relation.

To briefly recall the dataset characteristics, it is comprised by a total of 269,648 images from the Flickr network, annotated with a total of 81 semantic categories. Each image has multiple tags and may belong to multiple semantic categories. We consider the 1000 more frequent tags for text representation [28, 131].

We extended this dataset to include timestamp information. Accordingly, we crawled images' metadata and stored the *datetaken* field to be used as timestamp. Images that are missing, do not have associated tags, or without timestamp are excluded. We only keep images from year 1999 to 2009[3], resulting in a 10 years corpus, with a total of 169,283 images. We use the NUS-WIDE dataset for temporal cross-modal embedding learning as some tags have been shown to reveal a dynamic behaviour [120]. Train, validation and test splits comprise 129,500, 22,854 and 17,112 instances, respectively. We use *years* as granularity for NUS-WIDE, with content spanning over a total of 11 years.

4.6.1.2 SocialStories Dataset

This dataset consists of a collection of social media documents covering a large number of sub-events about two distinct major events of interest for the general public. We

[3]The dataset was released in 2009, with its distribution having a mean of 2006.69 ± 1.175.

Table 4.2: List of SocialStories categories for EdFest2016 and TDF2016.

EdFest2016	*Audience/Crowd*, *Castle*, *Selfies/Group Photos/Posing*, *Fireworks*, *Music*, *Streets of Edinburgh*, *Food*, *Dance/Dancing*, *Show/Performance*, *Building(s)/Monuments*, *Sky/Clouds*, *Person*, *Water*
TDF2016	*Spectators*, *Bicycle/Pedaling*, *Road*, *Yellow-Jersey*, *Car/Truck*, *Peloton*, *Crash*, *Field(s)/ Mountain(s)*, *Buildings/Monument(s)*, *Food*, *Sky/Clouds*, *Water*, *Person*

created this dataset to fill a gap in the literature, w.r.t. cross-modal learning datasets from dynamic corpora.

In particular, we considered Twitter as a source of social media dynamic content. We specifically considered events that span over multiple days and that contain considerable amounts of diverse visual material. These are expected to have strong temporal correlation across modalities with respect to its semantics. Taking the aforementioned aspects into account, we selected the following events:

Edinburgh Festival 2016 [4] **(EdFest 2016)** - Consists of a celebration of the performing arts, gathering dance, opera, music and theater performers from all over the world. The event takes place in Edinburgh, Scotland and has a duration of 3 weeks in August. The dataset contains 82,348 documents. A total of 1,186 were annotated with 13 semantic categories (listed in table 4.2).

Le Tour de France 2016 [5] **(TDF 2016)** - Consists of one of the main road cycling race competitions. The event takes place in France (day 1-8, 11-17, 20-23), Spain (day 9), Andorra (day 9-11), Switzerland (day 17-19), and has a duration of 23 days in July. The dataset contains 325,074 documents. A Total of 747 were annotated with 13 semantic categories (listed in table 4.2).

We crawled content from Twitter, for the two aforementioned events. A set of manually selected and highly representative seeds (e.g. #TDF2016, #edfest2016) were used to collect tweets. Table 4.1 summarizes the dataset characteristics. After crawling content with event specific hashtags and seeds, we applied a set of content filtering techniques [11, 81] to discard SPAM and annotated documents event-specific semantic categories. Annotators were asked to annotate media documents (image and text) with one or more categories. An additional *None* category is shown, when none of the categories apply to the instance. We obtained a total of 1186 and 747 annotated pairs, with an average of 3.0 ± 1.47 and 2.4 ± 1.26 categories per instance, for EdFest2016 and TDF2016, respectively. For both events, we use 90% of the data for development and the remaining for testing. We further split the development data using 15% for validation. Accordingly, training, validation and test splits, for EdFest2016 are 906, 571 and 119, respectively, and

TDF2016 are 571, 101 and 75, respectively. We consider *days* as the temporal granularity for SocialStories, the content spans from 202 and 219 days for EdFest2016 and TDF2016, respectively.

4.6.2 Methodology

We evaluate the temporal cross-modal embedding model on the task of cross-modal retrieval. Namely, we consider two tasks: 1) *Image-to-Text* retrieval ($I \mapsto T$) and 2) *Text-to-Image* ($T \mapsto I$) retrieval.

Retrieval performance using mean Average Precision (*mAP@K*), which is the standard evaluation metric for cross-modal retrieval [28, 99, 121, 123, 133] and normalized Discounted Cumulative Gain (*nDCG@K*). We follow [28, 121] and set $K = 50$. For *mAP@K*, an instance is relevant if it shares at least one category. For *nDCG@K*, relevance is defined as the number of common categories. We complement our evaluation with a qualitative analysis.

We evaluate our temporal cross-model embedding model, TempXNet, with the three devised temporal correlations. Namely, we evaluate recency-based temporal correlations, **TempXNet-Rec** (section 4.4.1), semantic category-based temporal correlations, **TempXNet-Cat** (section 4.4.2), and latent-topic word-based temporal correlations, **TempXNet-Lat** (section 4.4.3).

We adopt as baselines the following methods:

- **CCA** [99] - Canonical Correlation Analysis, a linear embedding learning approach;
- **Bi-AE** [88] - A deep bi-modal Autoencoder;
- **Bi-DBN** [113] - An autoencoder of Deep Belief Networks;
- **Corr-AE** [28], **Corr-Cross-AE** [28] and **Corr-Full-AE** [28] - Different variants of the deep Correspondence Autoencoder;
- **DCCA** [2, 133] - Deep Canonical Correlation Analysis, a neural network-based extension to the CCA algorithm.

All baselines are atemporal and are described in 2.2.2.

4.6.3 Training and Implementation Details

Networks are jointly trained using SGD, with 0.9 momentum, and a learning rate of $\eta = 5 \times 10^{-3}$, with a decay of 1×10^{-6}. Weights $\boldsymbol{\theta}_V$ and $\boldsymbol{\theta}_T$, of the projection functions f_V

and f_T, are updated according the following update rule:

$$\theta_V = \theta_V - \eta \frac{1}{b} \nabla_{\theta_V} (\mathscr{L} + \lambda \cdot \mathscr{L}_{temp}) \qquad \theta_T = \theta_T - \eta \frac{1}{b} \nabla_{\theta_T} (\mathscr{L} + \lambda \cdot \mathscr{L}_{temp}). \tag{4.18}$$

Early stopping is used to avoid overfitting. Mini-batch size b is set to 10,000, and 64, for NUS-WIDE and SocialStories, respectively, and the total number of epochs is set to 25. For each neuron, we use *tanh* non-linearities. In SocialStories, DLDA was trained on the full un-annotated dataset. The number of latent topics P is set to 10. We set $D = 100$, $\lambda = 1.0$, triplet ranking loss margin $m = 1.0$, and Recency Bandwitdh $h = 0.3$, KDE bandwidth $h = 1$. We adopt the same image and text original features for TempXNet and baselines. Namely, we use TF-IDF bag-of-words representation for texts and activations of the penultimate layer of a pre-trained ResNet-50 CNN, on the ImageNet Large Scale Visual Recognition Challenge.

4.7 Experiments and Results

4.7.1 Cross-Modal Retrieval

We start by evaluating the proposed temporal cross-modal embedding learning model, TempXNet, with each of the three distinct temporal correlations on the task of cross-modal retrieval.

All methods are evaluated on the three datasets, of varying dimensions, representing corpora with different topic broadness, and thus distinct temporal dynamics. Table 4.3, Table 4.4 and Table 4.5, show the *mAP*@50 and *nDCG*@50 results for the NUS-WIDE, EdFest2016 and TDF2016 datasets, respectively.

The first observation we draw from the results is that TempXNet is highly effective across the three datasets, outperforming all the baselines, on both tasks, on all metrics. Specifically, TempXNet is able to rank at the top (*nDCG*) highly relevant instances (i.e. instances that share more semantic categories). This confirms our hypothesis regarding modeling temporal correlations, through temporal embedding learning.

Regarding the different temporal smoothing functions, in the NUS-WIDE dataset (table 4.3), distinct temporal correlations achieved identical performance. This confirms our hypothesis regarding the fact that the NUS-WIDE collection is not comprised by *dynamic* content, but is more like a photo collection where documents are related (*e.g.* belonging to the same event). However, for EdFest2016 and TDF2016 this is no longer the case, and performance oscillates. This suggests the existence of distinct temporal distributions, underlying each of these two datasets.

TempXNet-Lat outperforms the other correlations on EdFest2016. As TempXNet-Lat

exploits temporal correlations at the word level, it is able to capture correlations between instances based on word's temporal behaviour. Additionally, on EdFest2016, TempXNet-Rec outperforms TempXNet-Cat. This indicates that for EdFest2016, latent-based and recency-based temporal correlations are more preferred, instead of category-based correlations. Therefore, given the fact that TempXNet-Lat achieved better performance, words temporal behaviour, for this particular dataset, helps discriminating instances. Such behavior is expected when there are sporadic sub-events, provoking shifts on word's usage. Namely, every day has different artists, shows, etc.

Table 4.3: Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on NUS-WIDE.

Method	$I \mapsto T$		$T \mapsto I$		Avg	
	mAP	*nDCG*	*mAP*	*nDCG*	*mAP*	*nDCG*
CCA [99]	74.2	84.4	68.7	80.7	71.5	82.6
Bi-AE [88]	74.1	84.9	69.1	80.0	71.6	82.4
Bi-DBN [113]	69.5	81.7	53.7	67.8	61.6	74.7
Corr-AE [28]	76.2	86.3	74.3	83.9	75.2	85.1
Corr-Cross-AE [28]	72.8	84.4	74.8	84.4	73.8	84.4
Corr-Full-AE [28]	75.4	86.0	75.5	84.6	75.5	85.3
DCCA [2, 133]	73.9	85.1	76.1	85.0	75.0	85.1
TempXNet-Rec	78.7	86.6	79.9	87.6	79.3	87.1
TempXNet-Cat	78.8	86.6	**80.0**	**87.7**	**79.4**	**87.2**
TempXNet-Lat	**79.1**	**86.9**	79.5	87.4	79.3	**87.2**

Table 4.4: Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on EdFest2016.

Method	$I \mapsto T$		$T \mapsto I$		Avg	
	mAP	*nDCG*	*mAP*	*nDCG*	*mAP*	*nDCG*
CCA [99]	58.6	75.5	53.3	73.7	56.0	74.6
Bi-AE [88]	64.9	83.8	66.4	83.0	65.7	83.4
Bi-DBN [113]	56,7	78.3	46.7	67.1	51.7	72.7
Corr-AE [28]	67.8	85.8	67.8	83.0	67.8	84.4
Corr-Cross-AE [28]	60.0	80.6	64.3	81.4	62.2	81.0
Corr-Full-AE [28]	68.0	85.4	68.7	83.2	68.3	84.3
DCCA [2, 133]	89.7	96.2	72.4	85.5	81.1	90.9
TempXNet-Rec	94.5	97.4	**95.5**	97.7	95.0	97.6
TempXNet-Cat	94.0	96.9	93.6	97.3	93.8	97.1
TempXNet-Lat	**96.4**	**98.6**	**95.5**	**98.1**	**96.0**	**98.4**

Table 4.5: Cross-modal retrieval results (*mAP*@50 and *nDCG*@50) on TDF2016.

Method	$I \mapsto T$		$T \mapsto I$		Avg	
	mAP	*nDCG*	*mAP*	*nDCG*	*mAP*	*nDCG*
CCA [99]	58.0	76.9	57.7	75.4	57.8	76.2
Bi-AE [88]	72.5	88.6	67.0	82.2	69.7	85.5
Bi-DBN [113]	64.5	82.9	56.1	74.2	60.3	78.6
Corr-AE [28]	73.5	89.1	71.4	86.1	72.4	87.6
Corr-Cross-AE [28]	70.5	85.9	72.2	86.3	71.4	86.0
Corr-Full-AE [28]	74.1	89.4	71.8	86.5	73.0	88.0
DCCA [2, 133]	88.4	95.5	73.8	86.2	81.1	90.9
TempXNet-Rec	87.2	93.9	89.1	94.6	88.2	94.3
TempXNet-Cat	**92.6**	**96.8**	**91.5**	**95.9**	**92.1**	**96.4**
TempXNet-Lat	88.1	94.7	90.3	95.8	89.2	95.3

The fact that TempXNet-Rec is the second best-performing approach on EdFest2016, hints that structuring data that is close in time, in the same neighborhood, yields an effective structuring of the embedding space for that collection.

On TDF2016 dataset, TempXNet-Cat outperforms all the other baselines and correlations by a considerable margin. This result indicates that for this dataset, focusing on semantic categories temporal density distributions helps achieving a better structure, and retrieving more relevant content. This may be due to the existence of distributions with multiple modes (e.g. periodic dynamic behavior). In fact, TDF2016 topics are to some extent periodic, e.g. stages, cyclists, mountain races, news regarding winners, etc.

In Figure 4.8 and Figure 4.9 we show the *mAP* results per category (the average of *mAP*@50 in the two directions $I \mapsto T$ and $T \mapsto I$), on the EdFest2016 and TDF2016, respectively. For EdFest2016 the top-performing correlation was the TempXNet-Lat. In Figure 4.8 we can see that it obtains better performance on most categories. However, for the *Fireworks* category, we can see that TempXNet-Rec achieves better performance. Fireworks happened once during the Edinburgh Festival, and by using distance in time to measure temporal correlations and structure data according to that type of correlation, yields better structuring. In TDF2016 the top-performing correlation was TempXNet-Cat. Again, from Figure 4.9, we can see that while in most categories it obtained better performance, in some categories the other correlations performed better. For instance, for the *Crash* category, both TempXNet-Rec and TempXNet-Lat outperformed TempXNet-Cat. The same happened for the category *Spectators* and *Person(s)*. Looking at the results on these datasets with dynamic corpora (EdFest2016 and TDF2016), we identify the following pattern: for categories covering topics that happened sporadically and/or at distinct and specific moments in time, w.r.t. the duration of each event, both TempXNet-Rec and TempXNet-Lat achieve better performance. Examples of such categories are *Fireworks* and *Crash*. On the other hand, for categories that cover topics that occur during the whole event (*e.g.* *Food* and *Road*), TempXNet-Cat achieves better performance.

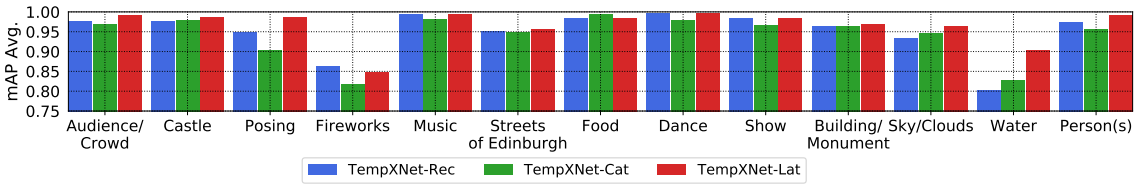


Figure 4.8: Cross-modal retrieval *mAP* results, average of $I \mapsto T$ and $T \mapsto I$, per category, on EdFest2016.

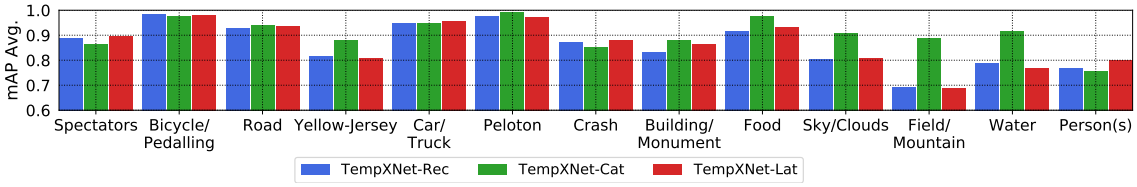


Figure 4.9: Cross-modal retrieval *mAP* results, average of $I \mapsto T$ and $T \mapsto I$, per category, on TDF2016.

Figures 4.10 and 4.11 shows the precision-scope curves for both EdFest2016 and TDF2016 datasets, respectively, on the Image-to-Text and Text-to-Image tasks. On the x axis we vary the value of k, and the y axis shows the corresponding $P@k$ (described in section 2.4). On EdFest2016, it can be observed that TempXNet-Lat always outperforms the remaining correlations. Similarly, on TDF2016 TempXNet-Cat also outperforms the remaining correlations, which is consistent with the previously discussed results. In both datasets, performance drops roughly linear across all methods.

In the presence of datasets with different intrinsic temporal distributions, our temporal cross-modal embedding learning model is able to effectively model such distributions, provided that a suitable temporal correlation is used. Apart from the three temporal correlations evaluated, TempXNet can accommodate any other temporal distributions.

4.7.2 Media temporal correlations

In this section we perform a qualitative analysis of the different temporal correlations. The goal is to assess how well temporal correlations are captured by each temporal model. To this end, we query each model and compare its relevant instances distribution with the true ground-truth temporal distribution. Specifically, we perform two queries, one for EdFest2016 in which the target are instances of the semantic category *Castle* and one for TDF2016 in which the target are instances of *Crash*, respectively. Each query comprises only the textual modality, corresponding to the $T \mapsto I$ setting. The top-50 retrieved results are evaluated, either relevant or not-relevant are considered. The two top performing temporal correlations (TempXNet-Cat and TempXNet-Lat) and the DCCA baseline are considered. For each query we show four sample images. Figure 4.12 depicts

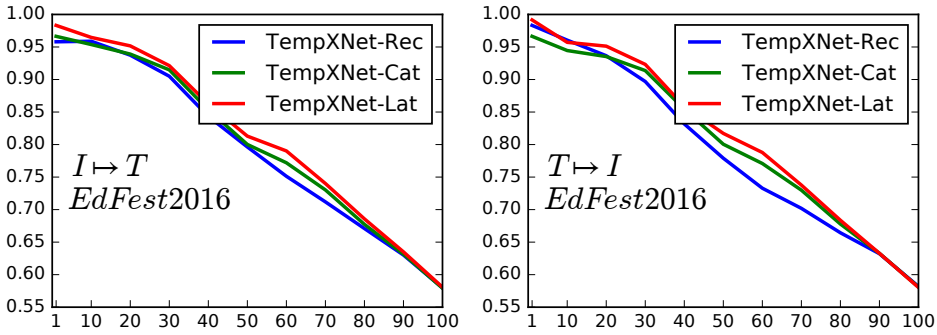


Figure 4.10: Precision-Scope curves for Edinburgh Festival 2016.

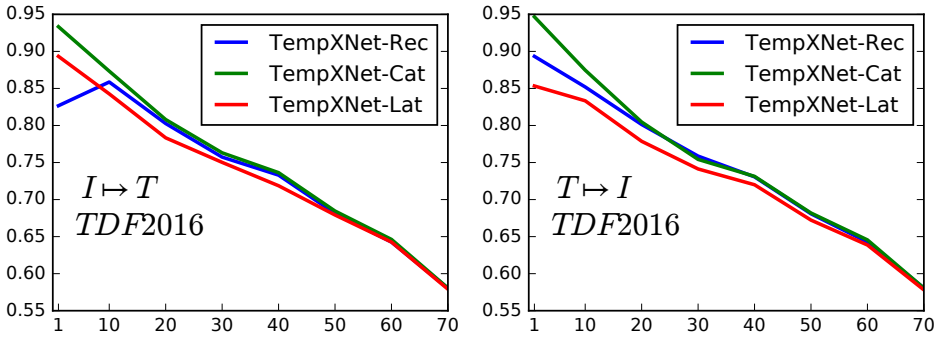


Figure 4.11: Precision-Scope curves for Tour-de-France 2016.

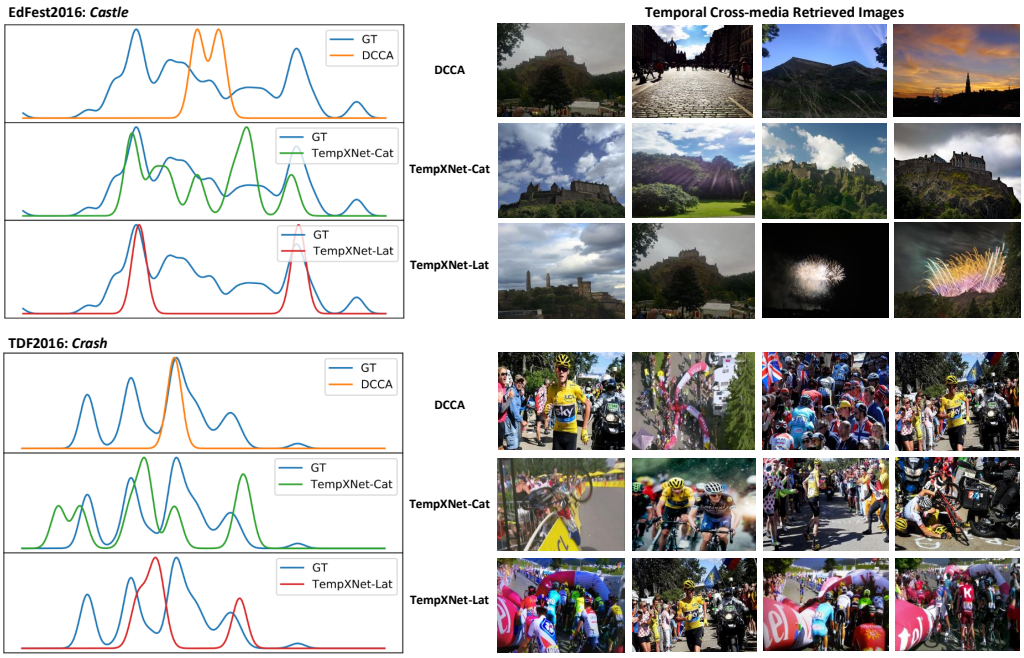


Figure 4.12: Qualitative analysis of the different temporal correlations on the EdFest2016 and TDF2016 dataset. Each plot depicts the temporal distribution of ground-truth instances, from the categories *Castle* and *Crash*. We use *days* as time granularity.

the result of this experiment. Each plot depicts the temporal distribution of ground-truth (GT) and relevant instances retrieved by each model, with the x-axis corresponding to time. We normalize each distribution to the $[0, 1]$ range (min-max normalization).

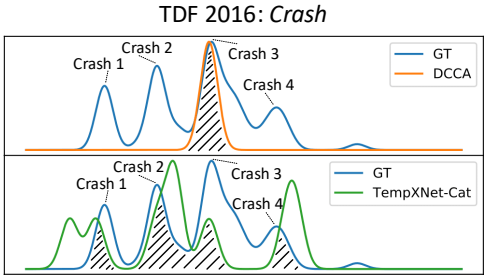


Figure 4.13: Temporal vs. Non-Temporal method.

On the EdFest2016 plot, one can observe that the temporal distribution of the semantic category *Castle* has multiple peaks, with the two larger ones being on the beginning and the end, respectively, of the dataset time span. These correspond to the beginning and ending of the festival, where at the beginning, there was a light show, and at the end, there was a fireworks show, both taking place at the Castle. Although the temporal correlations are different, it can be seen that both TempXNet-Cat and TempXNet-Lat are able to cover both larger peaks, by retrieving relevant instances at the corresponding moments in time. Even though TempXNet-Cat achieved a better fit to the ground-truth, TempXNet-Lat achieved better retrieval results. This may be due to the fact that it covers the most salient peaks.

On TDF2016 there were several crashes during the event, and this is reflected in the peaks of the ground-truth. Given the somewhat periodic nature of these peaks, TempXNet-Cat reveals a better fit to the GT curve. The fit of the TempXNet-Lat correlation is slightly worse, as it is based on individual word dynamics, and despite the periodic shape of the category *Crash*, words that occur in *Crash* instances may not reveal this pattern (e.g. usually it refers to racers names, etc.). The DCCA baseline completely fails to capture the temporal distribution of relevant documents. Given these observations, we verify that our model can effectively grasp temporal correlations of data, and structure the embedding accordingly.

To complement the illustrations we presented and discussed in section 4.1 regarding the modeling of several craches that took place in Tour de France 2016, we isolate and annotate the previous figure, to compare the best performing temporal model (TempXNet-Cat) to the best performing atemporal model (DCCA). The result is shown on figure 4.13. By inspecting the results we can see that TempXNet-Cat is capable of retrieving content from all craches, unlike the atemporal model, and achieve a better fit to the ground-truth curve. This evidences the capability of temporal cross-modal embeddings to structure data according to both their semantic and temporal correlations, and that this structuring contributes to the model performance.

4.8 Critical Summary

In this chapter we looked into the important problem of modeling semantically similar media that vary over time. Current state-of-the-art cross-modal methods assume that collections are static, overlooking visual and textual correlations (and cross-correlations) that change over time. TempXNet evaluation exposed the four fundamental concluding points:

- **Temporal cross-modal embedding**. This was the first work to propose time in cross-modal embedding learning. It derives from the idea that multimedia data should be organized according to their semantic category and temporal correlations across different modalities. Several key components make the creation of this embedding possible.
- **Principled temporal soft-constraints**. The creation of the embedding is temporally constrained by estimating temporal correlations of semantic categories and words, encoding the underlying dynamics of modalities. The investigated forms of soft-constraints stem from well-grounded statistical principles leading to a solid and rigorous optimization framework. Hence, modality projections are learned through neural networks, coupled by the same loss function, subject to the aforementioned temporal constraints.
- **Models of temporal cross-modal correlations**. We observed that temporal correlations are seldomly simple as the recency model of temporal correlations was never the best model. In fact, we could contrast the results in the EdFest2016 and the TDF2016 datasets and conclude that both datasets follow different distributions: the EdFest2016 has several one time shows and events, and the TDF2016 contains several repeated events.
- **Improved retrieval precision in dynamic domains**. Accounting for temporal cross-modal correlations improved cross-modality retrieval across all datasets. The proposed TempXNet models outperformed past cross-modal models. Moreover, the best retrieval precision was obtained by the TempXNet-Cat and TempXNet-Lat that model temporal correlations with different levels of granularity.

Chapter 5

Diachronic Cross-modal Embeddings

In this chapter we investigate embeddings that retain the temporal dimension of data. In the previous chapter, we formulated the TempXNet model, which uses time information in a *relative manner*, to structure data in a *static* temporal embedding space. The model structures visual and textual elements according to their temporal and semantic correlations. A function f_t, that quantifies temporal correlation between any two elements (either visual or textual) is used to decide how close should two elements be embedded in a common space. It turns out that while relative temporal information is used to organize the embedding space, after the learning phase, the **temporal dimension is not preserved**, thus any information regarding visual-textual correlations evolution is lost.

Instead, in this chapter we seek for models where the temporal dimension is explicitly incorporated in the model in an absolute manner. Therefore, these embeddings will be able to **capture the evolution of visual-textual correlations over time**. By doing so, we enable a set of novel multimedia understanding operations, that are supported by the diachronic embeddings. These will be demonstrated in section 5.5.

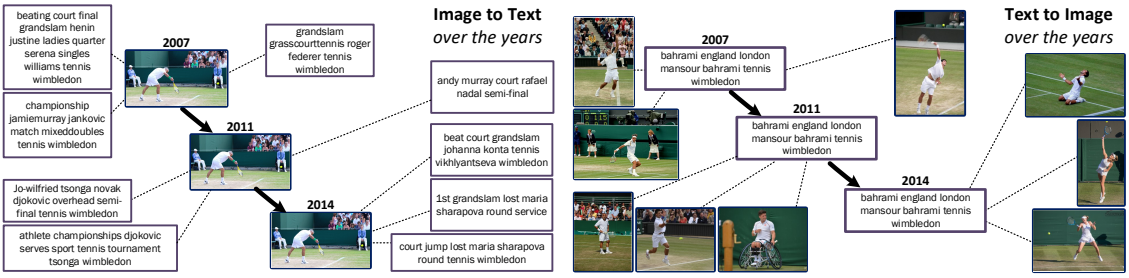


Figure 5.1: Diachronic Cross-modal Embeddings illustration.

5.1 Formulating the Diachronic Embedding Space Hypothesis

While visual elements can be seen as anchors – *i.e.* pictures and videos freeze a reality and do not change afterwards – word usage and meaning do change over time (*e.g.* as discussed above, the same image, or a semantically similar one, may be referred twice, at different points in time, but with different descriptions). This requires the development of a continuous model, in which **time is explicitly modeled, thus allowing conditioning on time at both training and inference time**.

Models with such capacity, are often referred as *Diachronic*[1] models. Accordingly, a diachronic model captures how something changes over time. Models with this expressiveness have been researched in the field of natural language processing for understanding language evolution. These capture language evolution, by analyzing words' usage over time [41, 101, 136]. For a more in-depth discussion of these models please refer to section 2.3.2.2. In the spirit of the nomenclature adopted in the NLP field, we coin embedding models that capture the evolution of the correlations between vision and language, as **Diachronic Cross-modal Embeddings**.

The cross-modal scenario brings new challenges, towards obtaining a diachronic model. For language, *diachronicity* stems from the evolution of words' co-occurrences, which is often referred as words' context. In the cross-modal scenario, *diachronicity* stems from the evolution of visual-textual correlations, i.e. interactions between vision and language. Therefore, a novel common embedding space must be sought where the space structure accommodates the temporal dimension, in order to capture data interactions over time. Figure 5.1 illustrates an application enabled by a diachronic cross-modal embedding, in modeling data temporal context. On the left part of the figure, we analyze the most related texts, for the same visual semantics, at different periods in time. On the right part of the figure, we analyze the most related images to the same text, over

[1]Definition of Diachronic: "Relating to the changes in something, especially a language, that happen over time", Cambridge dictionary.

time.

The rationale is that cross-modal interactions evolve along the temporal dimension. Therefore, the embedding space should structure images and texts such that for each instant t, elements are organized according to semantic correlations between instances and their corresponding absolute timestamp. This results in a model in which neighbors of an element (e.g. a text or an image), at time instant t_1, may differ from the neighbors at time instant t_2, if data correlations between the two instants change (Figure 5.1). In practice, this corresponds to learning projection functions $f_V(\cdot)$ and $f_T(\cdot)$, that giving an image or text, respectively, and a temporal instant, project the instance in a continuous cross-modal embedding. Projected instances should then **lye close to instances with the same semantic and temporal context**.

To achieve this, two main challenges should be explicitly addressed:

a) **unveiling** and **quantifying**, for each image and text, the evolution of the semantic correlations w.r.t. to other instances;

b) **Designing a model and optimization objective** capable of learning the diachronic embedding, in which at each instant t, semantic category information is used to guide the structuring of multimodal instances' neighbourhood.

To tackle these two challenges we employ a two-part approach for neighborhood structuring for an arbitrary instant time t: **first**, for instances of the same semantic category within a given time range, semantic correlations need to be maximal, **second**, instances outside a given time range are placed far apart. Then, an adapted triplet ranking loss is formulated to achieve a continuous diachronic structure. This is achieved by enforcing correlations from the two dimensions (semantic + temporal) in the space structure organization. The temporal context of instances is considered, at each instant, to align instances on adjacent time instants.

It should be noted that unlike in the previous chapter, where the Temporal Cross-modal Embedding requires the adoption of a specific temporal distribution, in a diachronic setting we avoid committing to a specific distribution. The goal is to preserve the time dimension, and consequently preserve all the original temporal traits of data. Therefore, the rationale is that given a corpus that contains multimodal data, computationally represented using heterogeneous representations, we want to obtain an embedding that unifies those representations and preserves original data timelines.

5.2 Embedding Definition

The key element of diachronic embedding methods is the set of time-preserving projection functions, responsible for mapping the original data onto the embedding. This section will detail the proposed embedding and the corresponding projection functions.

5.2.1 Diachronic Cross-modal Space

We start by recapping the notation introduced in section 1.1.3, and defining the task of diachronic cross-modal embedding learning. Without loss of generality, let $C = \{d_i\}_{i=1}^{N}$ be a set of N *visual-textual* instance tuples

$$d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, ts^i, c^i), \tag{5.1}$$

where $\mathbf{x}_V^i \in \mathbb{R}^{D_V}$ and $\mathbf{x}_T^i \in \mathbb{R}^{D_T}$ are the feature representations of the image and textual elements, respectively, ts^i the timestamp and c^i the instance (unique) semantic category. Accordingly, D_V and D_T correspond to the image and text features dimensionality, respectively. The instances timestamp have a timespan defined by $TS = [t_{start}, t_{end}]$, where t_s and t_f are the first and last instants of the dataset, respectively.

The goal is to obtain a common continuous (over time instants) embedding space, in which the visual and textual elements are organized according to their semantic category and timestamp. The diachronic space is formally defined as follows:

Definition of Diachronic Cross-modal Embedding

Definition 2. A *diachronic cross-modal embedding space* refers to a common space, that structures visual and textual elements of data instances over time. In this embedding, similarity between instances of the same category that are close in time, is maximized ($s(x_*^1, x_*^2) \rightarrow 1$). In all other cases, similarity between instances is minimal ($s(x_*^1, x_*^2) \rightarrow 0$).

In diachronic word embeddings, the temporal dimension captures *word meaning change* [41, 136]. This contrasts with the cross-modal scenario in which words (textual modality) that co-occur with pictures (visual modality), at distinct time instants, may change. This characteristic is quantified by the semantic alignment of diachronic models (Definition 3), which will be defined and detailed in the next section.

5.2.2 Temporal Embeddings Alignment

Solving the two main challenges introduced in section 5.1, requires tackling the core problem of modeling cross-modal data evolution, which is not present in diachronic word embeddings. Namely, when the projection unit is a *word*, it is assumed that each word occurs at least once, in each possible time instant [9, 40, 41, 101, 136]. This means that one can just worry about modeling the evolution of its co-occurrences, while knowing that it will be present at each instant. In the cross-modal scenario this assumption is not valid. The reasons are twofold:

1. An image is likely to be posted on a single or very few time instants. However, the visual concepts and the semantics depicted by the image, are expected to be referred several times. In a real scenario, every possible visual concept exists on all time instants, regardless of the existence of posts depicting those concepts;

2. Each image has descriptions with multiple words, that are likely to be different between semantically similar images.

Based on this, what can be assumed instead, is that each visual/textual concept is present at every time instant, but the associations of each visual concept with textual concepts may evolve. Therefore, the two modalities, visual and textual, should be *aligned* based on their semantics under the rationale that correlation should be retained, at adjacent time instants. This leads us to the formulation of the *temporal alignment problem*, which is essential to address the challenges presented in the previous section:

Definition of Temporal Alignment problem

Definition 3. The *Temporal Alignment problem* states that in a cross-modal diachronic embedding space, correlation between instances x_*^1 and x_*^2 of the same category that are close in time, according to a given temporal window w, should be retained, i.e. $s(x_*^1, x_*^2) \rightarrow 1$.

In this definition, category information is used as a proxy for semantic similarity, to align instances based on their semantics.

Solving this problem, while learning the target embedding, will ensure local smoothness on adjacent time instants, for semantically similar instances. As pointed out by [9], which considers this issue as one of the main aspects to learn a diachronic word embedding, this is crucial to make sure that our model will capture trajectories, w.r.t. to similarity within instances, that reflect cross-modal semantic drifts.

5.2.3 Time-preserving Projections

The diachronic cross-modal embedding embeds the modality vector $\mathbf{x}_*^i$ (image or text) of an instance d^i and a time instant ts, using the pair of functions:

$$\mathbf{e}_V = f_V(\mathbf{x}_V^i, ts; \boldsymbol{\theta}_V, \boldsymbol{\theta}_{time}) \quad \mathbf{e}_T = f_T(\mathbf{x}_T^i, ts; \boldsymbol{\theta}_T, \boldsymbol{\theta}_{time}), \tag{5.2}$$

where each $\mathbf{e}_* \in \mathbb{R}^D$ denotes the embedding of the corresponding modality, at time instant t. $\boldsymbol{\theta}_V = [\boldsymbol{\theta}_{V_h}; \boldsymbol{\theta}_{V_o}]$ and $\boldsymbol{\theta}_T = [\boldsymbol{\theta}_{T_h}; \boldsymbol{\theta}_{T_o}]$ are the model parameters, and $\boldsymbol{\theta}_{time}$ are the key parameters responsible for controlling the temporal structuring of the $\mathbf{x}_V^i$ and $\mathbf{x}_T^i$ projections. As a consequence, parameters $\boldsymbol{\theta}_{time}$ are shared by the two projection functions.

Diachronic cross-modal embedding functions are defined by the mappings:

$$f_V(\cdot) : \mathbb{R}^{D_V} \times TS \mapsto \mathbb{R}^D \qquad f_T(\cdot) : \mathbb{R}^{D_T} \times TS \mapsto \mathbb{R}^D. \tag{5.3}$$

The output of f_V and f_T is normalized such that $||f_*(\cdot)||_2 = 1$. Accordingly, instances will be organized based on time and semantic similarity, over a D-dimensional hypersphere. Similarity between projected sample elements $\mathbf{x}_*^i$ and $\mathbf{x}_*^j$, is computed through *cosine* similarity, namely by the dot product $s(\mathbf{x}_*^i, \mathbf{x}_*^j) = f_*(\mathbf{x}_*^i) \cdot f_*(\mathbf{x}_*^j)$. The resulting embeddings, produced by f_V and f_T will be characterized by the properties defined in the following section.

5.2.4 Embedding Properties

The structure of the temporal embedding space, *i.e.* how multimodal instances will be organized, is formalized by a set of fundamental properties. These properties stem from two grounding intuitions: data is primarily associated by the temporal dimension, and then by their semantic categories. The model will thus capture the evolution of semantic correlations, over time instants, by maximizing the similarity between instances that are within a given temporal window and share the same category. Formally, this is established by the following properties:

- **Property 1**. Two embedding vectors $\mathbf{e}_*^i$ and $\mathbf{e}_*^j$ will be projected into the same neighborhood, if the timestamps of d^i and d^j are within the same temporal window, *i.e.* $|t^i - t^j| \leq w$, and the two instances d^i and d^j share the same category, *i.e.* $c^i = c^j$.

- **Property 2**. Two embedding vectors $\mathbf{e}_*^i$ and $\mathbf{e}_*^j$ will be projected onto different neighborhoods, if the timestamps of d^i and d^j are outside the same temporal window, *i.e.* $|t^i - t^j| > w$, independently of the instances' semantic category;

- **Property 3**. Two embedding vectors $\mathbf{e}_*^i$ and $\mathbf{e}_*^j$ will be projected onto distant regions if the two elements do not share any semantic category.

The final and most novel property follows from the requirement that the target embedding space needs to be continuous over time. Thus, a final *semantic alignment over time* property is introduced:

- **Property 4**. For each image or text of an instance d^i, embeddings evolve smoothly between neighboring time instants t^1 and t^2, with $|t^1 - t^2| \leq w$. Formally, for x_*^1 and x_*^2, where $|t^1 - t^2| \leq w$, we have that $s(x_*^1, x_*^2) < \epsilon$.

The aim is to have a constant ϵ proportional to the distance in time ($|t^1 - t^2|$) between two instances. As will be detailed in section 5.3.4.1, the model definition is designed to achieve this behavior.

To make the four properties consistent, w is the same constant value across all properties. The window size should be defined based on the granularity in which time is being modeled and on the type of structuring one wants to achieve. In practice, we aim to obtain a generic diachronic model, and define w as a small value ($w = 4$).

Having defined the properties underlying the diachronic cross-modal space, we will now introduce the neural model and detail how these properties are materialized.

5.3 Diachronic Embedding Model Design and Learning

To learn the time-dependent continuous embedding functions $f_V(\cdot)$ and $f_T(\cdot)$ we define the optimization objective and show how it enforces the temporal organization of the embedding.

5.3.1 From Projections to Triplet Ranking Loss

Following the definition of the diachronic projection functions from section 5.2.3, a two component correlation scheme (temporal and semantic) is employed, where these components are encoded in the properties defined in section 5.2.4.

Building on the most recent state-of-the-art cross-modal learning works [92, 107, 121, 124, 135], we adopt the *triplet ranking loss function* as the model base loss. In its general formulation, triplets (x_*^a, x_*^p, x_*^n), are composed by an anchor element x_*^a, that should be more similar to positive elements x_*^p sharing a category, than to negative elements x_*^n not sharing categories, by at least a margin m. Triplet constraints are expressed as $s(x_*^a, x_*^p) > s(x_*^a, x_*^n) + m$, and then turned into a differentiable function, by means of a

relaxation under the hinge loss function [46]:

$$\ell_\theta(x_*^a, x_*^p, x_*^n; \boldsymbol{\theta}) = [m - s(x_*^a, x_*^p) + s(x_*^a, x_*^n)]_+, \tag{5.4}$$

where m denotes a constant margin, $[x]_+$ the function $max(0, x)$, and $\boldsymbol{\theta} = [\boldsymbol{\theta}_V; \boldsymbol{\theta}_T; \boldsymbol{\theta}_{time}]$ is the complete set of parameters. One of such constraints would then be enforced for each sampled triplet. In the next section we detail how triplet ranking loss is adapted to cope with the temporal dimension.

5.3.2 Joint Diachronic Triplet Ranking Loss

The learning problem is then formulated by coupling the learning of the two individual modality (f_V and f_T) and a third timestamp embedding function (f_{time}), through a global loss function $\mathscr{L}$. The full loss function of our model, for diachronic cross-modal embedding learning, is derived by enforcing multiple constraints, for each possible anchor element, and summing all the constraint violations. To this end, we define each f_* as a neural network, capable of unveiling complex non-linear interactions. The objective function becomes

$$\underset{\boldsymbol{\theta}_V, \boldsymbol{\theta}_T, \boldsymbol{\theta}_{time}}{\arg\min}\ \mathscr{L}(C; \boldsymbol{\theta}_V, \boldsymbol{\theta}_T, \boldsymbol{\theta}_{time}), \tag{5.5}$$

with $\boldsymbol{\theta}_V$, $\boldsymbol{\theta}_T$ and $\boldsymbol{\theta}_{time}$ being the projection function parameters.

State-of-the-art cross-modal retrieval interlace modalities by enforcing triplet constraints in both modality directions [85, 121, 124], *i.e.* $image \mapsto text$ and $text \mapsto image$. Thus, we formulate the final loss $\mathscr{L}$ function for diachronic cross-modal embedding model with parameters $\boldsymbol{\theta} = [\boldsymbol{\theta}_V; \boldsymbol{\theta}_T; \boldsymbol{\theta}_{time}]$ as:

$$\mathscr{L}(C) = \sum_{a,p,n} \underbrace{\mathscr{L}_\theta(x_V^a, x_T^p, x_T^n; \boldsymbol{\theta})}_{image \mapsto text} + \underbrace{\mathscr{L}(x_T^a, x_V^p, x_V^n; \boldsymbol{\theta})}_{text \mapsto image}, \tag{5.6}$$

where p and n denote indices of positive and negative instances, respectively, w.r.t. an anchor element x_*^a. This function is evaluated batch-wise. Thus, at each batch, the sampled elements are used to create triplet constraints.

5.3.3 Binned Structure

The most simple way to achieve an embedding model which preserves the time dimension, consists of applying data binning. In fact, most diachronic word embedding models apply this strategy [41, 65]. In this extreme case, data can be first divided into bins and a static cross-modal embedding model, optimized using the objective function from eq. 5.5,

where $\mathscr{L}_\theta$ is the standard triplet loss implemented by eq. 5.4, is trained on data from each bin.

The main issue is that as embeddings are obtained from a stochastic method, *i.e.* a neural network with its weights randomly initialized, embedding spaces from different bins will be **incompatible**. This incompatibility stems from the fact that different geometric space organization are expected to be achieved.

Even though correlations may shift across two adjacent bins, we can assume that these will change smoothly [41, 65] (**Property 4**). Therefore, we can perform embedding alignment within each adjacent bins. However, as previously discussed in section 5.1, in diachronic word embedding learning methods, it is assumed that a word will appear in all instants, and therefore in all bins. This enables one to use a strategy that focuses on aligning the embeddings of each word, at each adjacent time instant (each bin). In our setting, this assumption is not possible as each instance is likely to occur only once. Instead, we assume that each possible visual/textual semantics, not the actual instances, are presented in every time instant.

We start by adapting the binned embedding model from diachronic word embeddings [41, 65] to the cross-modal scenario. First, we align embeddings of adjacent bins by solving the Orthogonal Procrustes problem [41, 65, 105]. Namely, given two embedding matrices $\boldsymbol{M}_t$ and $\boldsymbol{M}_{t+1}$, each with shape $2 \cdot N \times D$, containing N images and N texts embeddings representative of adjacent time instants, the best rotational alignment is computed as:

$$\Omega_{t \mapsto t+1} = \underset{\Omega^T \Omega = \boldsymbol{I}}{\arg\min} \|\boldsymbol{M}_t \Omega - \boldsymbol{M}_{t+1}\|_F, \tag{5.7}$$

preserving cosine similarities within each $\boldsymbol{M}_t$. The problem is solved through single-value decomposition [105] (SVD). Matrix Ω is an orthogonal matrix, thus, it will only rotate/align the embeddings of $\boldsymbol{M}_t$ without changing their norm. The Procrustes alignment problem assumes that each ith entries of matrices $\boldsymbol{M}_t$ and $\boldsymbol{M}_{t+1}$, corresponds to the same object, and thus should be aligned. Our new assumption states that if a concept exists in bin t then it should also exist in bin $t+1$, and so on. Therefore, we set $\boldsymbol{M}_{t+1}$ by projecting images and texts from instant t in bin $t+1$, and then perform alignment. This means that instances from the previous bin are projected in bin $t+1$, using the corresponding projection functions. With this strategy, we ensure that the ith embedding, on both $\boldsymbol{M}_t$ and $\boldsymbol{M}_{t+1}$, corresponds to the same object. Algorithm 2 details all the steps of the algorithm.

After getting the set of rotation matrices Ω (one for each bin), embeddings $\boldsymbol{M}_{t+1}$ of bin $t+1$ are aligned by applying the inverse rotation (transpose), i.e. by the multiplication:

$$\boldsymbol{M}_0 = \boldsymbol{M}_0 \cdot \boldsymbol{I}, \qquad \boldsymbol{M}_{t+1} = \boldsymbol{M}_{t+1} \cdot \Omega_{t \mapsto t+1}^T. \tag{5.8}$$

Algorithm 2 Pseudocode for Binned Structure Embedding method.

Initialization: Corpus $C = \{d^i\}_{i=1}^{N}$ of N multimodal instances, with $d^i = (\mathbf{x}_V^i, \mathbf{x}_T^i, ts^i, c_i)$, over the timespan $TS = [t_{start}, t_{end}]$;
Hyperparameters: g binning granularity (e.g. months), embedding dimensionality D;

1: $BINS \leftarrow$ Bin instants on the interval TS, based on granularity g;
2: $num\_bins \leftarrow |BINS|$;
3: $M \leftarrow \{\}$;
4: $\Omega \leftarrow \{\}$;
{Get cross-modal embeddings matrices for each bin.}
5: **for** $t \leftarrow 1$ to num_bins **do**
6: $C_t \leftarrow$ Create corpus of bin t, by getting all instances d^i from bin t;
7: $f_V,\ f_T \leftarrow$ Learn a static cross-modal over instances of C_t, and obtain the modality projection functions;
8: $\boldsymbol{M}_{visual} \leftarrow$ Project and stack all images of C_t using f_V, resulting in a $N \times D$ matrix;
9: $\boldsymbol{M}_{textual} \leftarrow$ Project and stack all texts of C_t using f_T, resulting in a $N \times D$ matrix;
10: $\boldsymbol{M}_t \leftarrow$ Concatenate row-wise both matrices $\boldsymbol{M}_{visual}$ and $\boldsymbol{M}_{textual}$, resulting in a $2 \cdot N \times D$ matrix;
11: $\boldsymbol{M} \leftarrow \boldsymbol{M} \cup \{\boldsymbol{M}_t\}$;
12: **end for**
{Align embeddings by solving the Orthogonal Procrustes problem.}
13: $\boldsymbol{M}_{curr} \leftarrow \boldsymbol{M}(0)$;
14: **for** $t \leftarrow 1$ to $num\_bins - 1$ **do**
15: $\Omega_{t \mapsto t+1} \leftarrow \arg\min_{\Omega^T \Omega = \boldsymbol{I}} \|\boldsymbol{M}_{curr} \cdot \Omega - \boldsymbol{M}(t+1)\|_F$;
16: $\Omega \leftarrow \Omega \cup \{\Omega_{t \mapsto t+1}\}$;
17: $\boldsymbol{M}_{curr} \leftarrow \boldsymbol{M}(t+1) \cdot \Omega_{t \mapsto t+1}^T$;
18: **end for**
19: **return** Projection functions for each bin, Alignment matrices $\boldsymbol{M}$.

A clear disadvantage of a binned embedding is that as data is isolated at the bin level, correlations between instances of different bins are not accounted. Moreover, alignment is done locally, *i.e.* only between each adjacent bins. Therefore, even though during the iteration (line 13 of algorithm 2) some correlations from previous bins will be propagated to subsequent bins, it is expect that most inter-bin correlations will be lost. This will get worse the more far apart the bins are. To sum up, this model does not achieve *diachronicity*.

Despite these disadvantages, and while being prone to embedding alignment issues, this method preserves **temporal locality biases by definition**, as cross-modal correlations over distinct bins are never considered. This means that the model will be very good at keeping data biases from each individual bin t.

5.3.4 Continuous Diachronic Structure

In order to overcome the issues of a binned diachronic structure, we formulate a neural method that is explicitly designed to achieve a continuous diachronic structure. To achieve this, we start by materializing the diachronic triplet ranking loss function from equation 5.5. The cross-modality aspect is already addressed by the joint triplet ranking loss from equation 5.6, which interleaves modalities. Now, a novel loss function, that enforces the embedding properties defined in section 5.2.4, will be detailed.

To formulate the loss function, we adopt a two-component loss: a) inter-category component (section 5.3.4.2), which will be used to enforce the aspects of the properties that are defined between instances of different categories, and b) intra-category component (section 5.3.4.1), which will focus on the aspects that should be enforced within instances that share a semantic category. Formally, for each anchor element x_*^a, of a triplet (x_*^a, x_*^p, x_*^n), we define the following loss function:

$$\mathcal{L}(x_*^a, x_*^p, x_*^n; \boldsymbol{\theta}) = \mathcal{L}_{inter}(x_*^a, x_*^n; \boldsymbol{\theta}) + \mathcal{L}_{intra}(x_*^a, x_*^p; \boldsymbol{\theta}), \tag{5.9}$$

where $\mathcal{L}_{inter}$ and $\mathcal{L}_{intra}$ are based on the triplet ranking loss function, enforcing inter-category and intra-category embedding related properties, respectively. Specifically, $\mathcal{L}_{inter}$ is defined as in eq. 5.6.

Both x_*^n and x_*^p correspond to sampled negative and positive images or texts, respectively. We follow the strategy from previous chapters and sample triplets directly from mini-batches, and enforce triplet constraints for all instances, making full use of the information contained in the mini-batch [110]. To recap, for each instance $\mathbf{x}_T^a$ on a batch, we create triplets between an anchor instance $\mathbf{x}_*^a$ and all the negative instances $\mathbf{x}_*^n$ in the batch. Then, we use as positive element, its modality counterpart, *i.e.* if the anchor is an image ($\mathbf{x}_V^a$), we use a negative text ($\mathbf{x}_T^n$), and if the anchor is a text ($\mathbf{x}_T^a$), we use as negative an image $\mathbf{x}_V^a$. At each epoch, all samples are *seen* by the network.

Given that neural networks are used to obtain non-linear projections, the model will be trained using a batch-wise stochastic strategy. This raises several issues w.r.t. to obtaining a continuous diachronic structure that will now be discussed.

5.3.4.1 Intra-category and Temporal Smoothing

The function $\mathcal{L}_{intra}$ is responsible for enforcing two aspects: 1) approximate (**Property 1**) or separate (**Property 2**) instances of the same category, that fall within the temporal window of size w, respectively, and 2) perform embedding alignment over the same time window w (*i.e.* **Property 4**).

The assumption is that given an image or a text, its embedding should change

smoothly between adjacent time instants. Smoothness of that change is captured by the size of the considered window (**Property 4**). Property 4 implies that for an anchor element x_*^a, a temporal window of size w should be used to enforce smoothness between temporal neighborhoods in the embedding, for images and texts of the *same* category as x_*^a.

One strategy that resembles our temporal window is found in word embedding models. Namely, the context-window formulation, which has been quite successful in learning word embeddings [84, 94] to capture word context. This approach cannot be directly used in our problem as due to the stochastic nature of batch creation, we cannot guarantee that for an instance instance d^i and a temporal window of size w, we will have at least one element per time instant in the interval $[ts^i - w, ts^i + w]$. Therefore, for diachronic cross-modal embeddings, and inspired in the context-window approach, we consider temporally adjacent multimodal instances instead of a sliding window over text.

To accomplish this, a temporal window of size w is considered. The rationale is that from **Property 1**, we want instances of the same category to be close. But from **Property 2** definition, embeddings of instances of the same category, should be far apart, if they are temporally far apart.

If we do not enforce any margin between positive instances, the optimal solution is when all instances of the same category are mapped to the same point, losing temporal evolution of semantic correlations. To overcome this, we employ a temporally decaying triplet ranking loss formulation, for instances of the same category. Namely, given an anchor element x_*^a and a positive sampled instance x_*^p, $\mathscr{L}_{intra}$ is defined as the following branch function:

$$\mathscr{L}_{intra}(x_*^a, x_*^p; \boldsymbol{\theta}) = \begin{cases} 0 & , |t^a - t^p| \leq w \\ \rho(t^a, t^p) \cdot \ell_{\boldsymbol{\theta}}(x_{V/T}^a, x_{T/V}^a, x_*^p; \boldsymbol{\theta}) & , \ otherwise \end{cases} \tag{5.10}$$

where $\rho(t^a, t^b) = 1 - exp(-|t^a - t^b| \cdot \lambda)$ is a temporal decaying function, and λ the decay rate. $\ell_{\boldsymbol{\theta}}$ is defined in eq. 5.4. When two positive instances are less than w instants far part, no margin is enforced between the two. Otherwise, a triplet constraint is enforced, weighted by a decaying function that exponentially decreases the importance of ℓ_θ, the closer in time two instances are.

Finally, it is important to observe that the triplet sampling for batch creation makes a stochastic approximation to the optimal set of triplets (mining the optimal triplets for each batch is computationally too expensive [106]). Hence, although for a given batch it is likely that a given anchor x_*^a with timestamp t^a, has no adjacent (in time) projections, that same anchor x_*^a will eventually occur in subsequent batches with the required instances. Thus, convergence guarantees are preserved, due to the stochastic approximation made

through the triplet sampling strategy.

5.3.4.2 Inter-category separation

According to **Property 3**, elements that do not share any semantic category should be far apart. Therefore, for inter-category alignment $\mathscr{L}_{inter}$, we want to structure the embedding space such that instances of different categories will be far apart, independently of their timestamp. To this end, we enforce a triplet loss constraint, with a large margin (*i.e.* $m = 1$) over such triplets. Formally, $\mathscr{L}_{inter}$ is defined as:

$$\mathscr{L}_{inter}(x_*^a, x_*^n; \boldsymbol{\theta}) = \ell_{\boldsymbol{\theta}}(x_*^a, x_*^{(p=a)}, x_*^n; \boldsymbol{\theta}), \tag{5.11}$$

where $\ell_{\boldsymbol{\theta}}$ is defined in eq. 5.4. If the anchor is an image (*i.e.* x_V^a), then the positive corresponds to a text (*i.e.* x_T^a), and vice-versa. This formulation achieves two goals: enforces the separation of positive from negative instances, by a margin m, and aligns the embeddings of the image with the embeddings of the text of the anchor instance, and vice-versa. This also maximizes correlation between modalities using images and texts that occur together, what also contributes to better capture intra-category semantic diversity.

5.3.4.3 Triplet Sampling for Continuous Instance Structuring

To let the loss function of equation 5.9 enforce the required properties, special attention needs to be given to triplets' temporal and semantic correlations. As discussed in section 2.2.5.3 triplets are created in a batch-wise manner. For every instance d^i of a batch, we use d^i as an anchor to create triplets with all the others elements of the batch. Specifically, for each negative instance (*i.e.* different category), w.r.t. to d^i, a triplet is created and used in $\mathscr{L}_{inter}$. For each positive instance, w.r.t. d^i, a triplet is created and used in $\mathscr{L}_{intra}$. This means that for a batch of size B, the number of triplets created is $B \times B$. Then, as at each epoch, the network sees all the training instances, we can ensure that triplet constraints will be enforced for each possible instance.

5.3.4.4 Continuous Neural Projection Functions

The diachronic projection functions $f_V(\cdot)$ and $f_T(\cdot)$ are implemented as a neural network with 2 fully connected layers. Figure 5.2 depicts the neural architecture. Formally, the diachronic projection functions are defined as

$$f_*(x_*^i, ts^i) = tanh\left(\boldsymbol{\theta}_{*_o} \cdot \left[f_{h_*}(x_*^i);\ f_{time}(ts^i)\right]\right), \tag{5.12}$$

$$f_{*_h}(x_*^i) = tanh(\boldsymbol{\theta}_{*_h} \cdot x_*^i), \qquad f_{time}(ts^i) = tanh(\boldsymbol{\theta}_{time} \cdot ts^i), \tag{5.13}$$

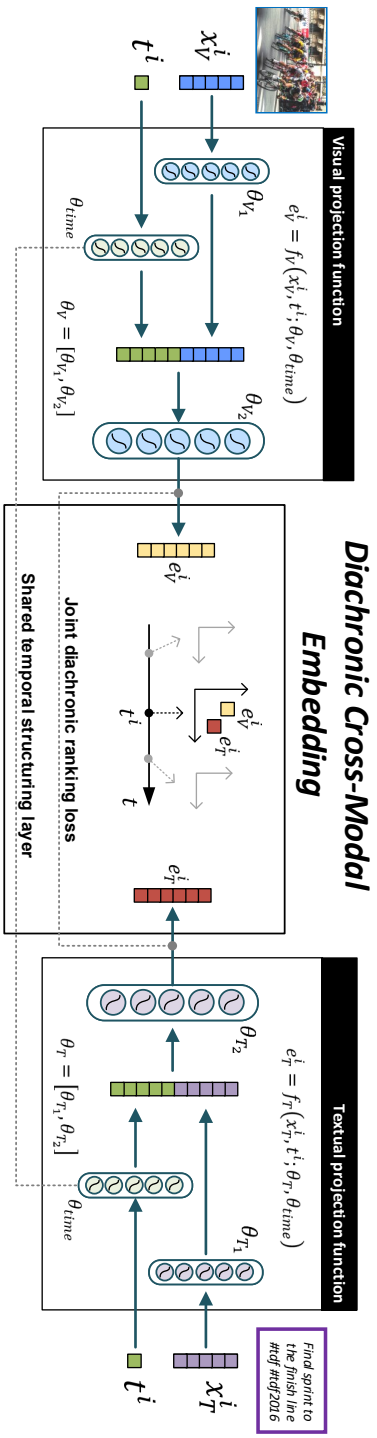


Figure 5.2: Diachronic cross-modal architecture overview. Visual (blue) and textual (purple) instances, at an instant ts^i, are mapped to a D dimensional diachronic embedding space. A shared temporal structuring layer takes the timestamp ts^i as input and learns an embedding for ts^i, that is then used to independently condition modality projections on time. A diachronic triplet ranking loss is responsible for structuring instances over time. Best viewed in color.

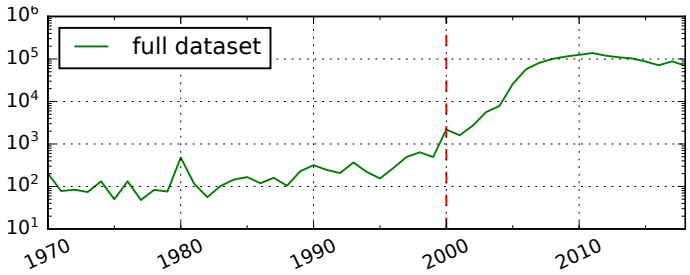


Figure 5.3: Temporal distribution of the full dataset. The x-axis shows the years while the y-axis shows the number of instances (log-scale). The red dashed vertical line delimits the cut performed due to low number of instances.

Table 5.1: List of categories of the 20 Years Flickr Images Dataset.

Categories	*Easter Sunday Edinburgh Festival Flood Formula One Horse Riding Independence Day London Marathon Mountain Camping Nuclear Disaster Olympic Games Picnic Rock Climbing Scuba Diving Snowboarding Solar Eclipse Terrorism Tour de France Tsunami The White House Wimbledon World Cup*

where $\boldsymbol{\theta}_{*_h}$, $\boldsymbol{\theta}_{time}$ and $\boldsymbol{\theta}_{*_o}$ correspond to hidden (per modality), time and output layer weight matrices, respectively. $[\cdot]$ denotes the concatenation operation. An initial **encoding layer** (f_{V_h} or f_{T_h}), receives the input vector and transforms it into an internal representation that is compatible with the internal representation of data timestamps. A **shared time embedding layer** (f_{time}) maps data timestamps to an embedding representation. The obtained time embedding is then used to condition the output projections of f_{V_h} and f_{T_h}, through a concatenation operation, making them time-dependent. A final **output layer** takes as input the result of conditioning f_{*_h} and f_{time} to produce the final D-dimensional projection to a diachronic embedding space.

5.4 Evaluation

In this section we evaluate the diachronic embedding model with a set of experiments. These are designed to evaluate two distinct aspects:

1. Evaluate the enforcement of the properties defined in section 5.2.4;

2. Exploit the embedding on novel multimedia understanding tasks, which allow the study of the evolution of multimedia information.

We start by describing the dataset in the next section 5.4.1 and the methodology in section 5.4.2.

5.4.1 Dataset - A 20 years Flickr Images Dataset

We are interested in obtaining a model that bridges vision and language over time. As such, and the ideal scenario is to apply our model to a dataset that mirrors reality, in the sense that it covers the events that took place in the last years, as well as the way that those events were captured (visually and textually), Therefore, we constructed a new large scale weakly-labeled dataset[2] with multimodal instances obtained from the Flickr[3] social network. This dataset fills a gap in the literature w.r.t. to large-scale, and large time span (years), multimodal datasets.

We collect documents related to topics that show a dynamic behavior over time such as spike-based, recurring and other type of events. Figure 5.3 shows the temporal distributions of eight sampled categories, and illustrates the diversity in terms of dynamic behaviour captured by the dataset. Data was collected over the period of 1-1-1970 to 31-12-2018. The Flickr API was used[4] to retrieve images and texts from a total of 21 categories, listed in table 5.1. We use the category name as keyword to query the API and collect data, and filter instances whose *date taken* is outside the considered temporal range. The topics within the 21 categories range from *periodic* major entertainment events (*e.g.* Edinburgh Festival, World Cup), natural disasters events (*e.g.* Floods, Tsunamis) which may have occurred more than once, but are not periodic, to more broad topics (*e.g.* Easter Sunday, Rock Climbing).

The models' granularity is set to *months*. To ensure that enough instances are available for each bin, we restrict the temporal range of images to the past 20 years (red line on figure 5.3 depicts the cut), and bins with less than 100 documents are excluded. After applying a set of SPAM filtering techniques, we obtain a total of 709,033 instances. In general, images have (near) professional quality. Texts are on average 23.0 words long. We use 10% of the data for testing and split the remaining data in 90% for training and 10% for validation, resulting in 574,308, 63,804 and 70,921 instances, for training, validation and testing, respectively.

5.4.2 Methodology

All experiments are performed in a cross-modal setting. Namely, we follow cross-modal embedding learning works [28, 92, 99, 121, 123, 133], and except stated otherwise, in each experiment we evaluate methods on the tasks of *Image-to-Text* ($I \mapsto T$) and *Text-to-Image* ($T \mapsto I$) retrieval, using mean Average Precision *mAP* as metric. For the experiments, only instances from the set split are used, with all images/texts being considered as

[2] https://novasearch.org/multimodal-diachronic-models/
[3] https://www.flickr.com/
[4] Only Creative Commons licensed data is retrieved.

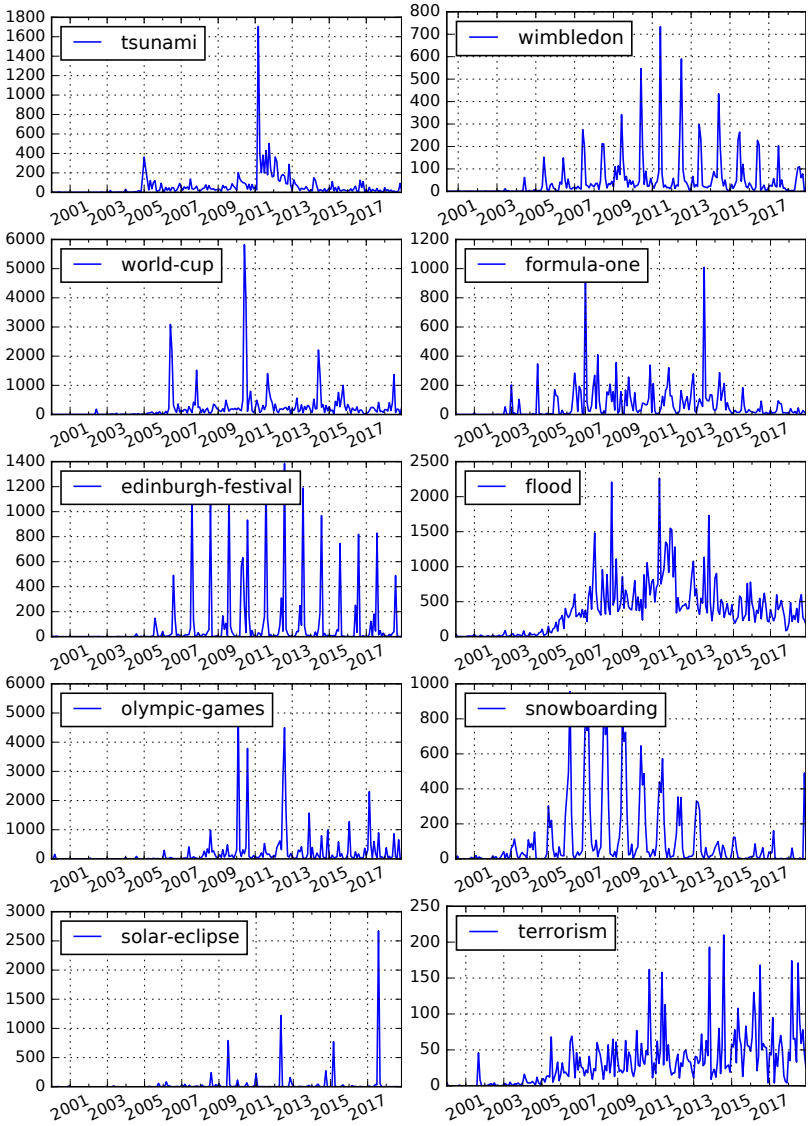


Figure 5.4: Temporal distribution of instances over eight sample categories. The dataset comprises content with high diversity in terms of temporal signatures.

queries. The training and validation split are only used to learn the models and for hyper-parameter tuning.

We refer to **DCM-Binned**and **DCM-Continuous** as the model with binned (section 5.3.3) and diachronic (section 5.3.4) structure. We recall that **DCM-Binned** is the method proposed by Hamilton et al. [41], but adapted to the cross-modal scenario. Both are trained using the full training split.

Additionally, we consider a **Static cross-modal** model, which consists of a cross-modal embedding in which time information is not used. This model shares all the architectural components of **DCM-Continuous**, except for the time embedding layer. The objective function is the standard triplet ranking loss. The training procedure, which is detailed in the next section 5.4.3, is the same for all models.

5.4.3 Training and Implementation Details

Networks are jointly trained using SGD, with 0.9 momentum, and a learning rate $\eta = 5 \times 10^{-3}$. We train the model for 25 epochs and retain the best performing model based on the validation set loss. Mini-batch size is set to 64. For each neuron, we use *tanh* non-linearities. A pre-trained ResNet-50 [45], with the last fully connected layer removed (softmax), is used for image representation. We set $\lambda = 0.1$, window size $w = 4$ and triplet ranking loss margin $m = 1.0$. We adopt the TF-IDF bag-of-words representation for texts and CNN image representations for all models.

The layers corresponding to the $\boldsymbol{\theta}_{*_h}$ and $\boldsymbol{\theta}_{time}$ parameters have dimension 1024 and 200 respectively, and $\boldsymbol{\theta}_{*_o}$ has $D = 200$ dimensions. Thus, for an instance d^i, the visual projection network takes the CNN representation $\mathbf{x}_V^i$ of the image, the textual projection a bag-of-words representation of the text $\mathbf{x}_T^i$, and the timestamp embedding the timestamp as input, producing the D-dimensional diachronic embedding.

5.5 Experiments and Results

In this section we define each of the conducted experiments, and present and discuss the results obtained. We conduct a set of experiments that assess the enforcement of the properties defined in 5.2.4. Moreover, across each experiment, we demonstrate the versatility of a diachronic cross-modal embedding, by showing how it can be used to tackle novel multimedia understanding tasks, in a principled manner. Namely, the diachronic embedding, enables the following media understanding tasks:

a) **Time Period based Inference**: given an image or a text, estimate its most relevant/likely time periods (section 5.5.1);

b) **Semantic Dispersion Understanding**: given an image or a text, understand its semantic dispersion w.r.t. to other instances, along the time dimension (section 5.5.2);

c) **Past and Future of Visual/Textual Concepts**: given an image or a text, project it onto the past or to the future, and understand how the concepts that it represents are manifested (section 5.5.3.2);

d) **Cross-modal Evolution Modeling**: given an image or a text, understand its trajectory w.r.t. to other images/texts (section 5.5.5).

These tasks are tackled by using a set of diachronic operations, supported by the diachronic cross-modal embedding. These operations will be described throughout the following sections.

Table 5.2: Media Time Period based Inference Results.

Methods (t-*mAP*@50)	$I \mapsto T$	$T \mapsto I$	Avg.
Static cross-modal	0.048	0.059	0.054
TempXNet-Rec (section 4)	0.052	0.070	0.061
DCM-Continuous	**0.126**	**0.144**	**0.135**

5.5.1 Time Period based Inference

The first experiment aims to **a)** empirically demonstrate the advantage of a diachronic cross-modal embedding versus a static cross-modal one, and **b)** evaluate the enforcement of the properties that define the neighborhood of each projected instance, based on a given time window w (**Property 1** and **Property 2**).

From the first two properties, it follows that two instances, of the same category, should be projected to the same neighborhood if their timestamps are within a given temporal window of size w (Property 1), otherwise they should be far apart. In this experiment we consider a very small time-window, *i.e.* $w = 1$. Accordingly, to evaluate this, we take each image and text as a query, and evaluate their neighbors in the diachronic cross-modal embedding. Specifically, we compute a *temporally bounded Mean Average Precision* (t-*mAP*): a neighbor d^j is considered relevant if it belongs to the same category and its timestamp is within a time-window of size w, w.r.t. to the timestamps ts^i and ts^j, of d^i and d^j, respectively. Formally, $|ts^i - ts^j| \leq w$.

To perform this experiment, we introduce the **Diachronic Operation #1**:

Definition of Diachronic Operation #1

Definition 4. Given an instance d^i as input, project it in its time instant ts^i, obtaining the embedding $\mathbf{e}^{i,ts^i}$, and compute its closest neighbors, while considering all the available instances. The output is a ranked list of neighbors d^j sorted in descending order of similarity $s(\mathbf{x}^i_*, \mathbf{x}^j_*)$.

By the DCM-Continuous, these will embed the instance close to semantically similar images and texts, that occurred near ts^i. Then, each image/text is projected in the embedding space, and its closest 50 neighbors (t-*mAP*@50) are evaluated.

We compare against a **static cross-modal** embedding, which discards time information, and against a temporal cross-modal embedding, learned with the TempXNet model, presented in chapter 4. For TempXnet, we use Recency as temporal correlation f_{corr}, which is the type of correlation that the DCM-Continuous model targets.

The results are shown in table 5.2. It can be observed that DCM-Continuous significantly outperforms the two compared models (twice the performance), in defining neighborhoods that respect both properties 1 and 2, *i.e.* that instances from the same category that are close in time (time window w), lye close together.

The TempXNet model performs better than the static cross-modal, but its performance is still highly inferior to the DCM-Continuous. The reason is that even though TempXNet uses time information to structure the space, it does so in a relative manner, and any information regarding how information is structured across time is lost. On the other hand, DCM-Continuous model learns a diachronic space which by definition retains information from the time dimension. Therefore, it manages to better structure instances such that they are close to instances of the same category, that occur at similar moments in time (Property 1 and 2).

In this experiment we infer the most likely time period for each image/text. Namely, given an image/text, we can analyze the closest neighbors in the diachronic embedding space, to assess the most relevant time periods for the corresponding image/text. Given that the diachronic cross-modal embedding jointly models visual, textual and time information, it offers a principled approach to tackle this task, that only requires a strategy to *inspect* the closest neighbors and select the most adequate time-periods.

5.5.2 Semantic Dispersion over Time

In this section we will examine the semantic dispersion of multimodal data over time. Given an image or a text, we expect that its correlations with other instances, over time, will evolve. The evolution pattern is expected to be grounded on the temporal characteristics (*e.g.* peak based, recurring event, etc.) of the topic of each instance.

Property 1 enforces similarity to be maximal, within instances of the same category and that fall within a given time-window. Additionally, **Property 4** states that the embedding of each instance d^i should evolve smoothly between neighboring time instants. For each instance, the model will individually enforce these two properties. However, if the instances representing the same semantics happen at distinct points in time, there may be the case that similarity between such two instances is high. One example would be on recurring events (*e.g.* snowboarding, Edinburgh festival, etc.). Such phenomena stems from the natural evolution of correlations between instances over time.

To assess this, we consider a set of target instances d^i, in which semantic evolution will be evidenced by semantic dispersion changes, over each instant $ts \in TS$. Namely, for an instance d^i with timestamp ts^i, given its embedding on instant ts^i, we define **semantic dispersion** as the variation of the similarity between d^i embedding and its closest neighbour instances, on a given time instant ts^j. Thus, to perform this experiment,

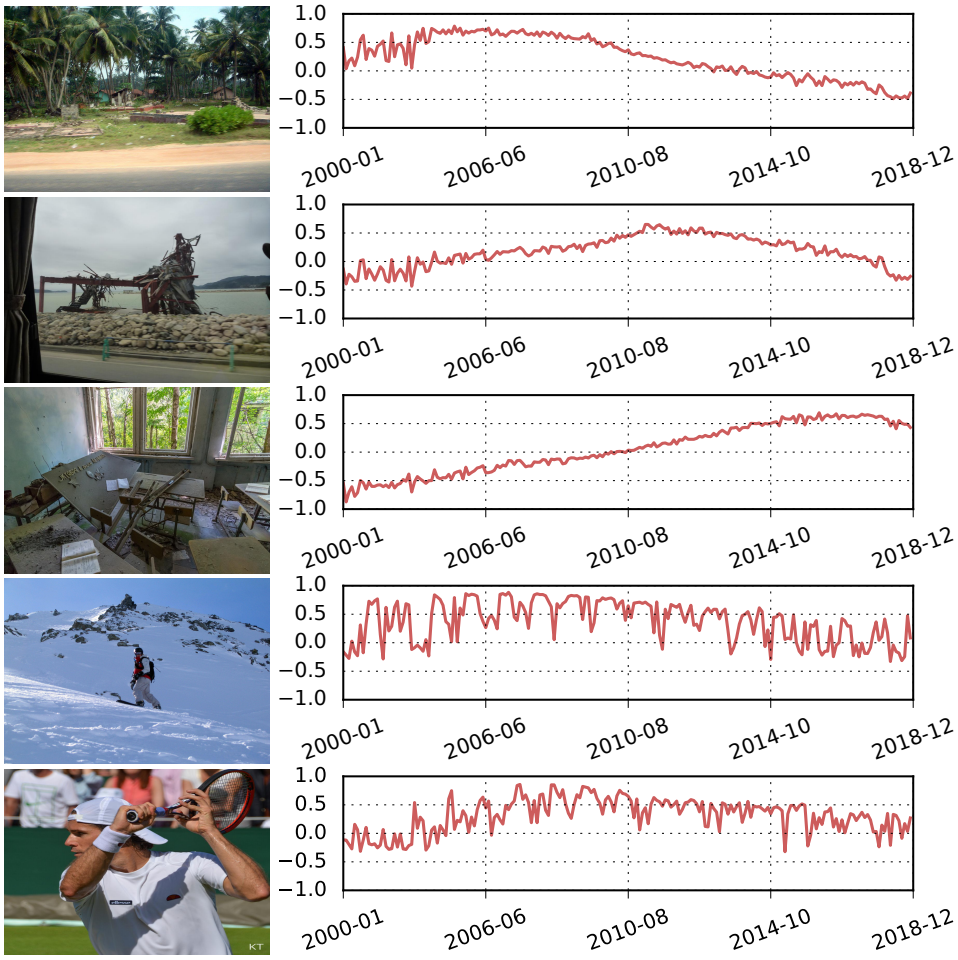


Figure 5.5: Semantic dispersion over time analysis for five sampled images. The y-axis denotes the similarity magnitude where 1 is maximal and -1 is minimal.

we introduce the **Diachronic Operation #2**:

Diachronic Operation #2

Definition 5. Given an instance d^i as input, project it in its time instant ts^i, obtaining the embedding $\mathbf{e}^{i,ts^i}$, and compute its closest neighbors, by considering solely as neighbors instances from a given time instant ts. The output is a ranked list of neighbors d^j, where $ts^j = ts$, sorted in descending order of similarity $s(\mathbf{x}^i_*, \mathbf{x}^j_*)$.

Accordingly, we sample a set of instances d^i and apply the diachronic operation #2. Therefore, each sampled d^i image is projected in its time instant ts^i (corresponding to its timestamp), using DCM-Continuous, obtaining the embedding $\mathbf{e}^{i,ts^i}$. Then, for each possible instant $ts \in TS$, we compute the semantic dispersion of the top K neighbours (texts), from that instant ts. Semantic dispersion on an instant ts is defined as the average of the cosine similarities between $\mathbf{e}^{i,ts^i}$ and each of the K neighbours on the instant ts.

We set $K = 5$.

Figure 5.5 shows the results of this experiment for five different images. The first two images belong to the *tsunami* category: the first corresponds to the Indonesia series of tsunamis in 2007, and the second to the tsunami in Japan, 2012. It can be seen that maximal similarity is achieved around the dates of the corresponding tsunamis. For the first image, after the peak around 2006, similarity decreases gradually in future instants. Despite the tsunami of 2012, its similarity with content from that tsunami is low. This evidences that DCM-Continuous effectively delivers Property 1 (section 5.2.4). The third image, taken in August 2018, depicts an abandoned place due to the *nuclear disaster* of Chernobyl (pictures of contaminated areas became possible with the advent of drones and robots). The first three images show that DCM-Continuous is able to deliver **Property 4**, by imposing a smooth evolution of instances embeddings.

The fourth and fifth images, *snowboarding* and *wimbledon tournament*, show a recurring evolution of semantic correlations over time. Semantic similarity over time in the fourth image gradually increases until 2015, and then stabilizes until 2018. As for the fifth image, similarity stabilizes between late 2006 and August of 2010, and then starts dropping. This experiment shows that diachronic embeddings obtained with DCM-Continuous encode cross-modal interactions evolution, enabling the understanding of multimodal correlations over time. Moreover, we verify that the model manages to preserve the original temporal traits of data.

In this experiment, we showed how we can use the diachronic cross-modal embedding to study media semantic evolution. Such analyzes is applicable to several scenarios ranging from historical analysis of events to trend detection, among others. Note that while we did this operation in an $I \mapsto T$ direction, the model also supports the $T \mapsto I$ direction, or even retrieving both images and texts.

5.5.3 Diachronic Semantic Alignment

In this section we evaluate the capability of both DCM-Binned [41] and DCM-Continuous, to capture and model diachronic data behaviours. First it is important to assess the semantic alignment over time of the diachronic space. This corresponds to assessing if the obtained diachronic embedding space is capable of relating embeddings of images and texts of instants ts^i, with instances that occurred in distinct time instants $ts^j \in TS$. As explained in the previous section, even though **Property 1** enforces minimal similarity between instances of the same category but happening at distinct points in time, due to the nature of content, similarity may still be retained on specific time instants (*e.g.* for recurring events). To accomplish this, we evaluate the semantic alignment quality of the diachronic space by designing two complementary tasks.

Table 5.3: Diachronic Semantic Alignment.

Coarse Semantic Alignment			
Methods (*mAP*)	$I \mapsto T$	$T \mapsto I$	Avg.
DCM-Binned w/ Align [41] (section 5.3.3)	0.203	0.197	0.200
DCM-Continuous	0.370	0.348	0.359
Local Semantic Alignment			
Methods (*mAP*@10)	$I \mapsto T$	$T \mapsto I$	Avg.
DCM-Binned w/ Align [41] (section 5.3.3)	0.078	0.086	0.082
DCM-Continuous	0.313	0.330	0.322

5.5.3.1 Coarse Semantic Alignment

The **Diachronic Operation #1** (introduced in section 5.5.1) can be used to understand which instances are semantically similar to an image or text, from the set of all instances, spanning across all possible time instants, based on their embedding similarity. The first type of semantic alignment uses this operation to understand how the embeddings of instances, projected in their original timestamps, correlate with the embeddings of instances from *all* instants (including distinct ones). Hence, the *coarse* designation.

An instance d^i is expected to be semantically correlated with instances not only on the d^i time instant, but also on other instants (*e.g.* recurring events). To capture such behavior, diachronic embedding models are required to correctly align embeddings over different time instants, such that semantic correlations are preserved. To do this, an image or text, should be embedded on the instant ts^i corresponding to its timestamp ts^i. Its neighborhood in embedding space, can then be analyzed by comparing the similarities of each d^i, against all projected instances, on their corresponding time instant (**Diachronic Operation #1**).

This is evaluated by projecting all images/texts d^i, from the test set, in their corresponding time instant ts^i (its timestamp), from which we obtain the embedding $\mathbf{e}^{i,ts^i}$. For each d^i, we then evaluate it against all the instances d^j on the test set, which are also projected in the embedding space in their timestamp, *i.e.* $\mathbf{e}^{j,ts^j}$. Accordingly, we use *mAP*, computed over the whole test set, to evaluate the semantic similarity of neighbors, using semantic category information.

The top part of Table 5.3 shows the results of this experiment. We observe that the DCM-Continuous significantly outperforms DCM-Binned, revealing superior *coarse* alignment capabilities. This is justified by the fact that in DCM-Binned, diachronicity is compromised due to the binned structure. Even though DCM-Binned uses a binning alignment procedure (detailed in section 5.3.3) it fails to align bins over distant time instants. The reasons are twofold:

a) Neural networks have a stochastic behavior, that stems from the network parameters initialization (we used Glorot [34], which is stochastic) and from the optimization strategy, which is the mini-batch stochastic gradient descent. Both aspects lead to different organization of data in the embedding space after convergence, despite its semantic correlations;

b) Data is binned, and each cross-modal embedding space from a bin t, is solely trained with data from that bin. While this isolation is a strong approach to preserve **temporal locality bias** (see section 5.5.4), it fails to capture correlations across time instants, and hence achieve a diachronic model.

The DCM-Continuous model manages to overcome these two aspects. The reason is that data is not binned, therefore, it is processed and modeled in a continuous manner, that enable capturing correlations across distinct time instants.

5.5.3.2 Local Semantic Alignment

The second type of semantic alignment is directly related to how the concepts of an image/text from an instance d^i, that took place in the time instant ts^i, are manifested (locally) at distinct time instants ts^j.

To assess this, we introduce the **Diachronic Operation #3**:

Diachronic Operation #3

Definition 6. Given an instance d^i as input, that occurred at instant ts^i, project it in a different time instant ts, obtaining the embedding $\mathbf{e}^{i,ts}$, and compute its closest neighbors, while considering only instances that occurred at ts. The output is a ranked list of neighbors d^j, such that $ts^j = ts$, sorted in descending order of similarity $s(\mathbf{x}^i_*, \mathbf{x}^j_*)$.

This operation is possible due to DCM's preservation of local alignment (w.r.t. to time). Namely, by **Property 1**, semantically similar instances should be close in the embedding space when projected into the same time instant ts^j.

Accordingly, we evaluate local semantic alignment by applying the **Diachronic Operation #3** and projecting instances d^i onto all possible timestamps $ts^j \in TS$ and assess how each projection $\mathbf{e}^{i,ts^j}$ relates to instances of that temporal neighborhood (each ts^j). For scalability reasons, we randomly sampled 50 query instances from each category, from the test set, and project each instance d^i into each timestamp $ts^j \in TS$. Then, for each time instant ts^j, we consider only the neighbors of the embedding of d^i on that instant, *i.e.* only embeddings $\mathbf{e}^{i,ts^j}$ of images and texts from ts^j instant are considered.

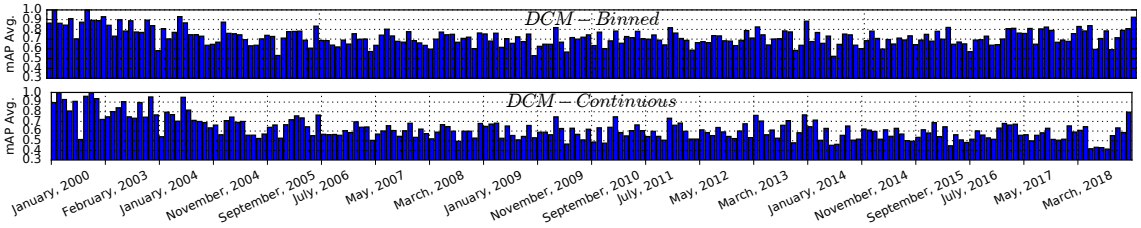


Figure 5.6: Temporally bounded cross-modal results (mAP) of DCM-Binned [41] and DCM-Continuous.

Table 5.4: Temporal locality bias preservation assessment.

Methods (*mAP*)	$I \mapsto T$	$T \mapsto I$	Avg.
Static Cross-modal (section 5.3.3)	0.646	0.631	0.639
TempXNet-Rec (section 4)	0.650	0.635	0.643
DCM-Binned w/ Align [41] (section 5.3.3)	**0.726**	**0.721**	**0.724**
DCM-Continuous (section 5.3.4)	0.632	0.616	0.624

Then we evaluate if the top-10 closest neighbours on each time instant are semantically similar (*i.e.* belong to the same category) using *mAP*@10.

Table 5.3 shows that DCM-Continuous clearly outperforms DCM-Binned. Again, this result follows the observations from the previous section 5.5.3.1, where DCM-Binned suffers from bad alignment across time instants, from which DCM-Continuous is able to overcome.

In this experiment, we showed how one can use a diachronic cross-modal embedding to study both the past and the future of an image or text, w.r.t. to how it correlates with other images and texts. Such task is supported by **Diachronic Operation #3**, which after being applied towards projecting an instance from an instant ts^i in an distinct instant ts^j, one can inspect the neighbors at the instant ts^j, or time instants close to ts^j.

5.5.4 Assessing the Preservation of Temporal Locality Biases

In this section we aim to assess the preservation of temporal locality biases. This refers to the inherent bias that exists in data from a specific time instant. As an example, the rationale is that even though the *Wimbledon tournament* occurs every year, if a model only sees data from a single year, it will be highly biased towards retaining cross-modal correlations, between images and texts, from that year.

As previously discussed, namely in section 5.5.3.1, binned models are expected to perform much better at retaining temporal locality biases as each bin embedding does not receive any influence from other bins. To confirm this, we evaluate the semantic organization of the embedding space in temporally-bounded data.

Namely, to perform this experiment, the test set is binned, and each bin is evaluated individually. Then, we introduce the **Diachronic Operation #4**:

Diachronic Operation #4

Definition 7. Given an instance d^i as input, project it in its time instant ts^i, obtaining the embedding $\mathbf{e}^{i,ts^i}$, and compute its closest neighbors, by considering solely as neighbors instances from its own time instant ts^i. The output is a ranked list of neighbors d^j, such that $ts^j = ts^i$, sorted in descending order of similarity $s(\mathbf{x}^i_*, \mathbf{x}^j_*)$.

After projecting all instances d^i and getting the embeddings $\mathbf{e}^{i,ts^i}$, we evaluate the closest neighbors, using *mAP*, and by considering all instances d^j, such that $ts^j = ts^i$.

Table 5.4 shows the results of the experiment. All DCM variants have shown to be on par with the remaining approaches, by obtaining scores above 0.60 *mAP* points. TempXNet-Rec, which uses time information to structure data in a static cross-modal space, outperforms a Static Cross-modal method, that completely ignores time. We observe that DCM-continuous obtains the lowest performance. The reason is that as we observed in sections 5.5.2 and 5.5.3.1, DCM-Continuous implicitly captures correlations across time instants, what negatively affects temporal locality biases.

This experiment confirms our hypothesis that semantic cross-modal correlations change over time. DCM-Binned outperformed all the other methods due to its capability to preserve temporal locality bias: for each bin (month) a static model is trained *solely* on data from that bin, thus modeling the local cross-modal correlations independently, and without influences from correlations of the remaining bins. However, it lacks the advantages of a fully diachronic model. Namely, the DCM-Continuous model is capable of retaining correlations across time instants, what is crucial to capture data evolution.

Figure 5.6 shows the plots of DCM variants, with the *mAP* results per month. It is clear that the DCM-Binned is superior across different months. This visualization also reveals that temporal locality biases are more present in certain months. Namely, in the first months, both methods achieve similar performances. In contrast, in the last months, DCM-Continuous degrades significantly. To conclude, if one wants a diachronic embedding that sacrifices the capture of cross-modal correlations evolution, to preserve temporal locality biases, the DCM-Binned model should be adopted.

5.5.5 Cross-modal Evolution

The cross-modal diachronic embedding model enables novel ways of exploring multimodal instances. One example is the analysis of the correlations evolution of an image

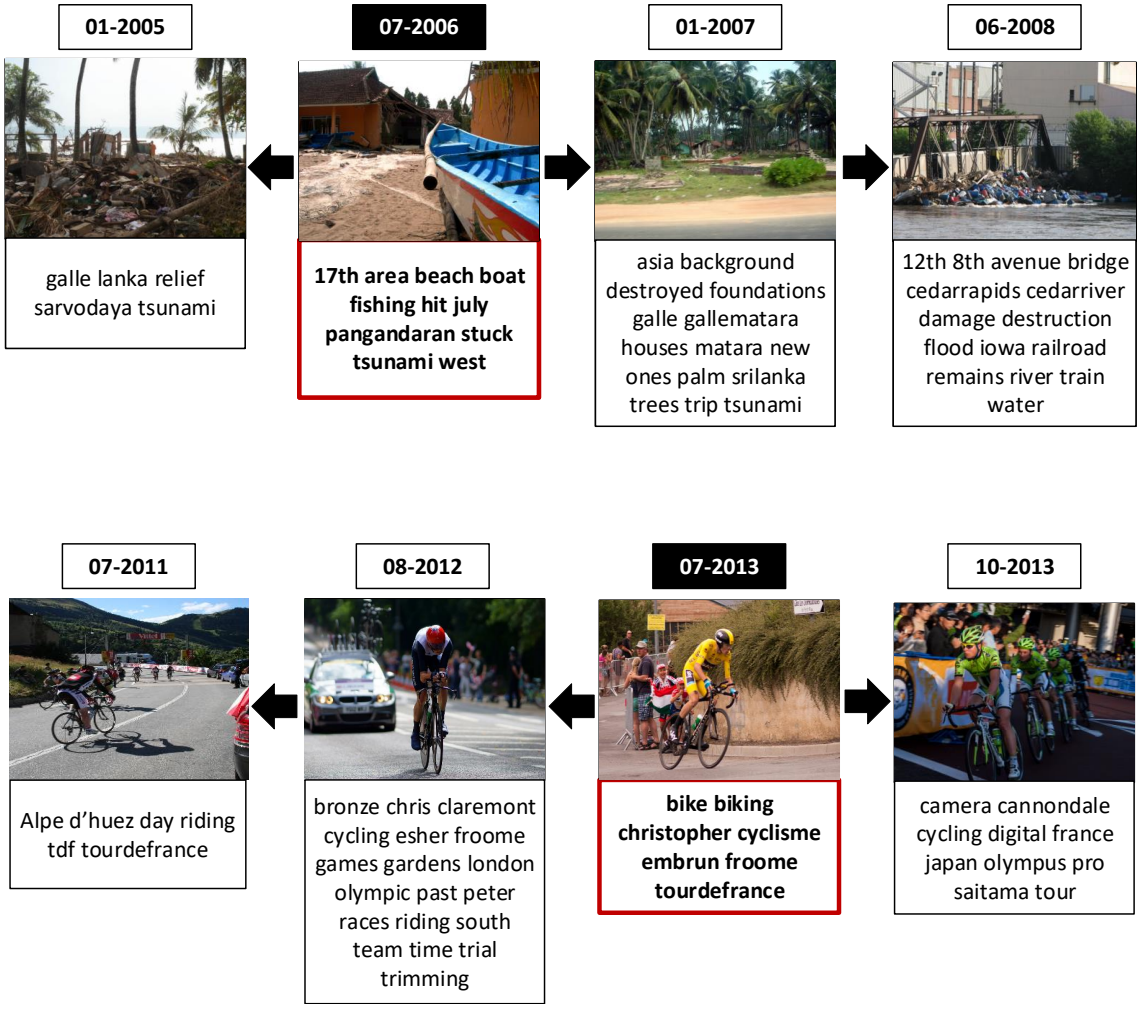


Figure 5.7: Evolution over time for 2 query examples (query timestamps are black-filled). Instances were retrieved from before and after the query timestamp. Image queries were used to retrieve documents through their text.

or text, along the years, which is encoded in its embedding trajectory. Such operation enables one to understand the correlations shift along the years.

To illustrate this, we perform an experiment in which we will use the diachronic cross-modal embedding to first, find the most relevant time instants for an instance, and then find the most relevant images/texts of these instants. Therefore, we start by randomly sampling a set of query images and texts, and projecting them on their corresponding timestamp instant. Namely, we first apply **Diachronic Operation #1** to get the embeddings $\mathbf{e}^{i,ts^i}$, of all the instances from the test set. Then, to avoid inspecting neighbors from all time instants, we restricted the number of instants to the top-20 ones, *i.e.* the bins with the closest image/text, to a target text/image, respectively, based on cosine similarity.

Figure ?? shows the results of the experiment, for two sample images (left part of the

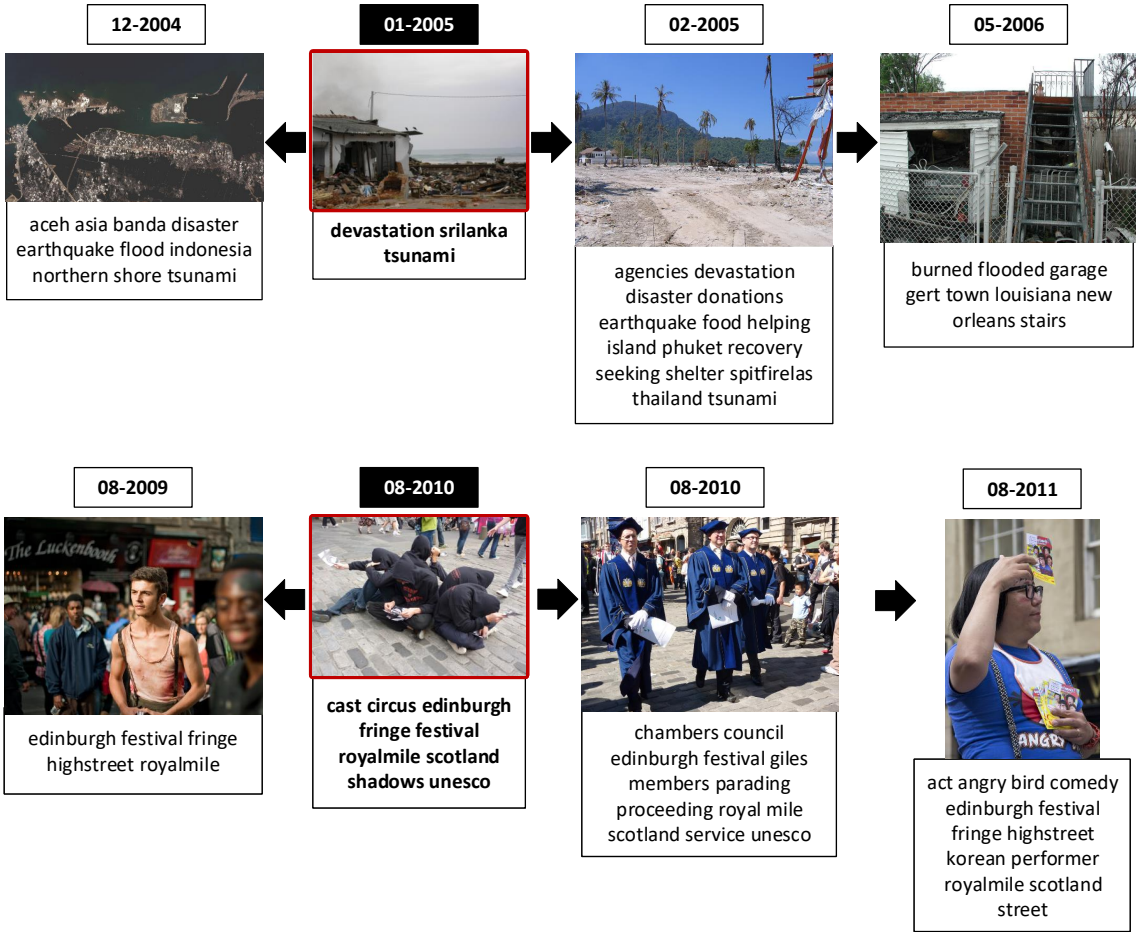


Figure 5.8: Evolution over time for 4 query examples (query timestamps are black-filled). Instances were retrieved from before and after the query timestamp. Text queries were used to retrieve documents through their images.

figure) and two sample texts (right part of the figure). Queries are marked with the black-filled timestamps and instances were retrieved from before and after this timestamp.

The inspection of the evolution timeline (obtained with DCM-Continuous) let us interpret the trajectory of the sampled images and texts, at particular time instants. In fact, this operation *automatically* reveals the stories of each image and text:

Indonesia Tsunami (Top-left Image) - the query is an image depicting the wreckage of the 2004 tsunami in Indonesia, from January 2005. When we navigate one month to the past, the closest neighbor text mentions the earthquake, that preceded the tsunami, and mentions floods that took place in the Banda Aceh city. By navigating one month to the future, the closest text talks about recovery and the food donations. By navigating further to the future, specifically 1 year and 3 months, the closest text reveals a similar disaster that occurred in May, 2006, in Louisiana;

Edinburgh Festival (Bottom-left Image) - the query is an image of a circus show, at the

Royal Mile street, during the 2010 Edinburgh Festival. Navigating either to the past or to the future, yields neighboring texts that both refer to the Edinburgh Festival. Note that even though the query was an image (the text was not used), DCM-Continuous manages to find texts that refer the Royal Mile street, which is mentioned in the text of the original query;

Sri Lanka Tsunami (Top-right Text) - the query is a text that refers a tsunami that occurred in Pangandaran, Indonesia, in 2006. In this example we navigate to the past and to the future roughly on yearly basis. We obtain an image of the wreckage of the tsunami that hit Sri Lanka in 2004 (the same as the one that hit Indonesia). When we inspect the future, we obtain an image from January, 2007, of the area that was affected by the 2004 tsunami. By looking one year further, we find an image of a related disaster, a flood that affected Iowa;

Tour de France (Bottom-left Text) - the query is a text from the Tour de France competition, in 2013, depicting the cyclist Christopher Froome at the Embrun stage. Going either to the past or to the future, yields images related to cycling.

To sum up, given an image or a text, with the diachronic cross-modal embedding one can automatically study long-term cross-modal correlations, across time, and find related content over different instants. This information is encoded in the trajectory of an instance, across time, which is modeled by the DCM-Continuous method.

5.6 Critical Summary

This chapter introduced the first diachronic cross-modal embedding, enabling novel interpretations of cross-modal semantic shifts over time. The key contributions of this chapter are:

- **Diachronic Cross-modal embedding**. The first Diachronic Cross-modal embedding learning approach, where the evolution of multimodal data correlations are modeled. Time is explicitly modeled, thus allowing *conditioning on time at both training and inference time*;
- **Time-dependent Neural Architecture**. The proposed neural architecture that can be conditioned on time. Namely, a novel **temporal structuring layer**, shared across the two projection functions, enables time-dependent projections.
- **Temporally constrained triplet ranking loss**. A novel triplet ranking loss formulation, that is temporally constrained, aligns instances embeddings over time, and enables the learning of neural projections from timestamped multimodal data;

- **Joint-inferences of image+text+time**. A principled approach that offers statistical guarantees, and allows for correct joint-inferences (image+text+time) that other methods do not, enabling it to be used for a wide number of media interpretation tasks.

Moreover, experiments, on a 20 year span dataset, illustrated the semantic evolution and temporal flexibility of the model. The key take-away messages are:

- **Cross-modal semantic evolution** is captured by the model, allowing the inspection of temporal multimodal information;
- **Time is handled in a flexible manner**, *i.e.* projections are timestamped, data is organised temporally, thus supporting several diachronical operations;
- **Exploitation of Long-term Data**. Static cross-modal methods need to be artificially fed with the relevant time period to achieve comparable results. Our model can infer the relevant time periods for each instance, by learning data evolution on long-term data.

Chapter 6

Conclusions and Future Work

The research conducted in this thesis aimed to investigate neural cross-modal embedding models that model interactions between vision and language over time. State-of-the-art cross-modal methods assume that collections are static, thus overlooking the evolution of visual and textual patterns of interaction. This thesis takes a step forward by bringing the time dimension to the core of approaches that focus on bridging vision and language. Accordingly, I investigated models that leverage on time information to structure multimodal data in a common cross-modal embedding space. This line of research focused on two complementary directions:

a) Investigate how to represent time information in cross-modal embedding models (Chapters 4 and 5);

b) Improve the expressiveness of cross-modal embedding learning methods, and consequently their effectiveness (Chapter 3).

6.1 Temporal Information on Cross-modal Embeddings

The main contribution of this thesis is on neural models that incorporate time information in cross-modal embeddings. Specifically, we proposed two distinct models for incorporating time: in a relative manner (TempXNet 4) and in an absolute or diachronic manner (DCM-Continuous 5).

I started by formulating the **Temporal Cross-modal Embedding**, which consists of a cross-modal neural embedding that considers time in a relative manner. This model

stemmed from the idea that **multimedia data from dynamic collections, should be structured according to their semantic and temporal correlations**. Given that in dynamic collections visual and textual correlations are constantly changing, it is important to consider the time dimension to unveil such evolution. As such, the proposed embedding uses temporal information to structure visual and textual data, according to their semantics *and* temporal correlations. First we identified the key components required to support the creation of such embedding space: **a)** estimate how much correlated in time (*i.e.* quantify) any two instances are, where different temporal distributions may be assumed to model data, and **b)** temporally constrain the embedding space according to the estimated temporal correlations, thus encoding the underlying dynamics of modalities in the embedding space.

The model was extensively evaluated on three distinct datasets: a standard benchmark dataset where the evolution of visual-textual correlations are expected to be minimal, and two datasets covering two distinct events, with highly dynamic data. The last two datasets were contributed to the community to support the evaluation of temporal cross-modal embeddings. From the results, I found that **incorporating temporal information can lead to better structuring of embedding spaces in dynamic datasets**. Also, we observed that when different temporal distributions are assumed and used to model data, we achieve different effectiveness. This result hints that different datasets potentially follow different distributions, and to achieve maximum performance the temporal distribution should be chosen according to the dataset at hand. The contributed model is formulated such that it is general enough to support different distributions, according to one's needs.

While *Temporal Cross-modal Embeddings* cannot bridge vision and language over time, it provided us the needed insights towards developing a more powerful model that tackles the thesis end-goal. Thus, with the work from chapter 4 we arrive at a diachronic formulation that overcomes the main limitation of the first approach: time is discarded after the learning stage. By overcoming this limitation, we arrive at the second major contribution of this thesis, the formulation of a **Diachronic Cross-modal Embedding**.

The **diachronic** model, proposed in Chapter 5, takes a step forward from previous approaches by being the first embedding that **models the evolution of patterns of interaction between vision and language**. The clear difference between the contributed model and previous work is that the model provides a principle approach that jointly models vision+language+time. The key contributions that made possible the diachronic model were: **a)** A two branch time-dependent neural architecture, in which each branch (corresponding to one modality) can be independently conditioned on time (input), and **b)** a novel temporally constrained triplet ranking loss formulation, that allows retaining the temporal dimension, structuring multimodal information across time, while aiming

to preserve the original data timelines.

The flexibility of the model was demonstrated in **supporting a set of multimedia understanding operations that require joint-inferences of visual+textual+temporal information**. The model capability to achieve a diachronic behavior, *i.e.* to capture the evolution of visual and textual interactions, was demonstrated experimentally. Moreover, experiments have shown how each of these operations can be used to extract rich insights from large multimodal datasets. In the spirit of multimodal machine learning approaches [8], which aim to leverage on the patterns of interactions between multiple modalities, and therefore combine information from multiple sources, diachronic cross-modal embeddings combines vision and language over time. Therefore, we were able to study the manifestation of specific concepts (materialized by an image or by a text), along the years (20 years span), while seamlessly relating visual and textual information.

To sum up, I showed that preserving the temporal dimension enables the study of multimodal semantic shifts on long-term data. Such feature is of critical importance to effectively grasp the temporal context of each image and text. Bringing time to cross-modal embeddings brings new ways to address multimedia understanding tasks that require framing content over time (*e.g.* multimodal retrieval, multimodal summarization, question-answering, etc.).

6.2 Learning of Neural Cross-modal Embedding Models

In a distinct but complementary research direction, we investigated alternative formulations to the cross-modal embedding learning framework. Namely, during the course of this research, I dived into the fundamentals of cross-modal embedding learning algorithms. This allowed me to comprehend and identify limitations in current state-of-the-art approaches. As a result, in this thesis I proposed an alternative formulation to one of the most widely used loss functions for embedding learning, the triplet loss. The aim was to provide it with more expressive power while still being aware of the optimization framework underlying neural models. Namely, by tying the enforcement of triplet constraints, in the embedding space structuring to the model optimization, towards exploiting the learning capabilities of neural models. To this end, I proposed an **Adaptive Maximum-margin Formulation**, that progressively activates the inference of pair-specific margins, towards enabling performing a fine-grain structuring of embedding spaces.

The proposed adaptive formulation was thoroughly evaluated on benchmark datasets. One of the findings was that by **augmenting the expressiveness of the most widely adopted loss function, triplet loss, better embedding structuring can be achieved**. With

an adaptive formulation, we achieved state-of-the-art performance on all datasets. We found that it is important to combine the benefits of large margins and then allow for small margins. While large margins are important to force separation of instances, imposing an high penalty when large margin triplet constraints are violated, small margins allow for fine-grain structuring by helping alleviating the constraint set's infeasibility problem and providing more informative errors to the model.

I believe that this novel formulation can have a significant impact across tasks that leverage on neural representation learning methods, making its adaptation and generalization to other tasks a promising research direction.

6.3 Limitations

In this section we discuss the limitations of the developed models. Namely, we go through each chapter and discuss the limitations and also some possible improvements.

6.3.1 Adaptive Margins

The adaptive margin formulation proposed in chapter 3 leads to effective cross-modal embeddings by increasing the expressiveness of the standard triplet loss. However, it has the following limitations and space for improvements:

- **Heuristic-based Adaptive Margin** - The adaptive margin is defined in an heuristic manner, making it task dependent. While the presented model offers the flexibility to define a custom adaptive margin formulation, making the adaptive margin function dependent on the model, *i.e.* on gradient information of each triplet, enables focusing on triplet constraints that add relevant information to update the model and makes the model general;

- **Parametric Scheduler** - The scheduler behavior is controlled by a set of parameters that must be tuned. Ideally the scheduler would also be tied to the gradient information from triplet constraints;

- **Hard-negative Triplet Mining** - The performance of the proposed method can possibly be improved by adopting an hard-negative triplet mining strategy [24, 106]. In fact, combining the triplet sampling strategy with the adaptive margin formulation should yield superior performance.

6.3.2 Temporal Cross-modal Embeddings

We proposed in chapter 4 the first temporal cross-modal embedding, which uses relative temporal correlation to structure data. During the design of the model, several decisions impose limitations on the model. Some of these limitations are:

- **Time is discarded** - The main limitation of this model is that time information is discarded after the training stage. This means that time information of new instances cannot be used for retrieval. In chapter 5 we present an approach that overcomes this limitation;
- **Fixed Granularity** - The model requires committing to a single temporal granularity (e.g. days, months, etc..). Depending on the granularity chosen, the temporal pairwise correlations captured will be different. Ideally the model would use a function for estimating temporal correlation that can cope with multiple granularity to make it more expressive. Devising a hierarchical model could allow for jointly capturing different granularity;
- **Temporal Distributions** - With the proposed model, to estimate temporal correlation, one must choose one temporal distribution. This assumption is highly broad, in the sense that it assumes that content from all categories, and all instances, are correlated in time based on the same temporal distribution. A possible extension to overcome this would be to have a method that based on instance information (*e.g.* category) would first choose the most adequate temporal distribution, and then estimate correlation.

6.3.3 Diachronic Cross-modal embeddings

While we have observed on the experimental evaluation that the Diachronic Cross-modal embedding model, proposed in chapter 5, is able to preserve *diachronicity* of vision and language, it still has some limitations that stem from its materialization:

- **Single granularity** - The model is only capable of modeling time on a single granularity. While different granularities can be support by controling the scale of the timestamps given as input, it does not support multiple granularities on the same diachronic embedding. Making the temporal granularity of the model *elastic*, would increase its diachronic expressiveness in terms of capturing temporal correlations at different time scales (e.g. from seconds to months);
- **Fixed time-window size** w - While a fixed time-window significantly simplifies the definition of the diachronic embedding, it also implies a commitment to the

value that is chosen for w, which has implications on the diachronic embedding structure. Namely, for some type of content it may make sense to have a small w while others may require a larger w;

- **Different temporal distribution shapes** - Content from different topics have distinct inherent temporal distribution shapes (Figure 5.4). Even though our diachronic model is designed to avoid committing to a specific temporal distribution, it may be necessary to account for these different distributions when structuring data (e.g. peak-based);

- **Supervised Setting** - The model was trained under a supervised setting. Namely, the topic categories are used as weak labels to build the triplet loss constraints. While this helps structuring data, provided that a general and large enough dataset is available, weak labels can be dropped and use image-text pair information to construct the triplets instead.

6.4 Future work

The research carried on this thesis aims to contribute to the end goal of devising human-like capable multimedia understanding systems. While the problem is far from being solved, this thesis opens two complementary research directions:

Computational Representations that Brige Vision and Language over Time - Embedding models that are highly effective and flexible, where time is taken as a first class citizen towards unveiling the semantics underlying vision and language interactions. Then, application of these models to different types of data collections such as medical repositories (*e.g.* Pubmed[1]), digital libraries and historical archives (*e.g.* WebArchive[2]), social media streams, and others, will enable the extraction of rich insights, encoded in the interaction between visual and textual information on these repositories;

Adaptive Neural Metric Learning Models - Adaptive optimization techniques that aim at achieving maximum expressiveness in defining how an embedding space should be structured, and consequently achieve better embedding space organization.

We intend to continue the research on these two directions. Thus, from each research direction, and from the findings on each chapter, stems the following promising research directions:

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://archive.org/web/

- **Generalization the adaptive triplet ranking loss**. - The proposed Adaptive Maximum-margin Formulation was applied to the learning of cross-modal embeddings. However, a large spectrum of tasks, such as Face Recognition [106], Visual Question Answering [1], Image Captioning [56], and methods, such as Generative Adversarial Networks [38], Memory Networks [114], and many more, heavily rely on triplet loss. I believe that the adaptive triplet-loss formulation can further improve these models by providing them more expressive power and therefore contribute towards addressing each of the aforementioned tasks;

- **Adaptive Temporal Cross-modal Embedding Learning** - In line with the previous direction, replacing the additive smoothing (eq. 4.7), that softly enforces the temporal constraints, by an adaptive margin formulation, in which temporal constraints are expressed in the margin function. These will enable providing the model much more accurate information about the mistakes that the network is doing, instead of a coarse-grain penalty;

- **End-to-end Adaptive Triplet Loss** - One of the main limitations of our adaptive triplet-loss formulation is that it requires parameter tuning. The next step would be to formulate the scheduler as a differential function that can be jointly optimized by gradient descent;

- **Temporal Granularities** - Both the Temporal and Diachronic Cross-modal embeddings were trained assuming a fixed temporal granularity (*e.g.* days, months). Being able to accommodate multiple granularities in the same embedding will be useful to model temporal correlations at different scales, in the same model, achieving richer data representations;

- **Unsupervised Diachronic Cross-modal Embedding** - The diachronic cross-modal embedding proposed in section 5 requires label information. Ideally the network architecture would be designed to be unsupervised (or self-supervised), enabling the use of a larger dataset, and thus obtaining a richer and more comprehensive embedding;

- **Multimodal Explanations for Long-term Data** - The diachronic cross-modal embedding captures the evolution of visual and textual information. The experiments performed had the aim of verifying the enforcement of the model properties and demonstrate the supported operations. Now, one can leverage on this model to study large collections of data, spanning several years (*e.g.* archives, long-term web data, etc.). Then one can use the model to understand the evolution of multimodal information: how events unfold (*e.g.* natural disasters, political events, and so on),

how did specific concepts changed over time, among many others. Eventually, event manifestations on the web would be characterized and systematized to a certain extent, using the proposed model, enabling us to understand the dynamics of web and user content contributions.

6.5 Forthcoming Challenges of Multimedia Understanding

Harming Artificial Intelligence systems with human-like cognitive capabilities, requires the capability of computationally understanding and reasoning about visual content, beyond the information comprised in the pixels. This requires framing visual, textual and context information jointly, to effectively capture visual content semantic context.

This thesis brought the temporal dimension to the modeling of visual and textual correlations. Namely, diachronic embeddings consist of a model that bridges vision and language over time. The challenge now is on making these representations completely general and expressive enough to capture the intricacies of the interplay between vision and language. This implies being able to model different modalities' interaction over time, without any explicit supervision. Whereas supervision helps guiding the structuring of embedding spaces with extra information, it may provide too coarse information that harms capturing of modalities' interactions.

Apart from temporal information, the context of an image or a text is also defined by other types of information. This can range from location (spatial information), user intents (*e.g.* mood, intention, induced emotions, etc.), to external *latent* causes (*e.g.* political scenario and orientation, crisis/war times, among others). Humans in general are highly skilled at perceiving all of these elements when interpreting a multimodal document.

To truly achieve systems that can understand multimedia information, all of this information needs to be cleverly combined. This requires the computational ability to not only combine but also jointly reasoning over visual, textual and all these contextual elements. In summary, **it requires both higher-order cognition and common sense reasoning about the world, on a particular context**. Towards this goal, some encouraging research work has been published recently, in bringing reasoning capabilities to visual question-answering systems [70, 140]. While the results are promising, more research is still needed to harm these kind of models with mechanisms to incorporate contextual information. The work on this thesis contributes to this long-term goal with neural representation learning models that can combine multiple modalities and also context information (time, specifically). The next steps point to a direction in which such neural embedding components are combined with reasoning models that can seamlessly

incorporate context information.

Bibliography

[1] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra. "VQA: Visual Question Answering." In: *International Journal of Computer Vision* 123.1 (May 2017), pp. 4–31. ISSN: 1573-1405. DOI: 10.1007/s11263-016-0966-6. URL: https://doi.org/10.1007/s11263-016-0966-6.

[2] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. "Deep Canonical Correlation Analysis." In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. ICML'13. Atlanta, GA, USA: JMLR.org, 2013, pp. III-1247–III-1255. URL: http://dl.acm.org/citation.cfm?id=3042817.3043076.

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. "VQA: Visual Question Answering." In: *International Conference on Computer Vision (ICCV)*. 2015.

[4] S. Asur and B. A. Huberman. "Predicting the Future with Social Media." In: *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*. WI-IAT '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 492–499. ISBN: 978-0-7695-4191-4. DOI: 10.1109/WI-IAT.2010.63. URL: http://dx.doi.org/10.1109/WI-IAT.2010.63.

[5] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, D. Joy, A. Delgado, A. Smeaton, Y. Graham, et al. "TRECVID 2018: Benchmarking video activity detection, video captioning and matching, video storytelling linking and video search." In: *Proceedings of TRECVID 2018*. NIST, USA. 2018.

[6] *Bag-of-Words Representation Illustration*. https://cdn-images-1.medium.com/max/1600/1*lnpjlG2auGBB4tvlEKTcgQ.png. Accessed: 2019-12-05.

[7] D. Bahdanau, K. Cho, and Y. Bengio. "Neural machine translation by jointly learning to align and translate." In: *ICLR 2015*. 2014.

[8] T. Baltrušaitis, C. Ahuja, and L. Morency. "Multimodal Machine Learning: A Survey and Taxonomy." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (Feb. 2019), pp. 423–443. DOI: 10.1109/TPAMI.2018.2798607.

[9] R. Bamler and S. Mandt. "Dynamic Word Embeddings." In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 380–389. URL: http://proceedings.mlr.press/v70/bamler17a.html.

[10] F. Barbieri, L. Marujo, P. Karuturi, W. Brendel, and H. Saggion. "Exploring Emoji Usage and Prediction Through a Temporal Variation Lens." In: *ArXiv* abs/1805.00731 (2018).

[11] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. "Detecting spammers on twitter." In: *Collaboration, electronic messaging, anti-abuse and spam conference (CEAS)*. Vol. 6. 2010, p. 12.

[12] Y. Bengio, P. Simard, and P. Frasconi. "Learning long-term dependencies with gradient descent is difficult." In: *IEEE Transactions on Neural Networks* 5.2 (Mar. 1994), pp. 157–166. ISSN: 1045-9227. DOI: 10.1109/72.279181.

[13] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin. "A Neural Probabilistic Language Model." In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 1137–1155. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944966.

[14] M. Benko. "Functional Data Analysis." In: *Statistical Methods for Biostatistics and Related Fields*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, pp. 305–327. ISBN: 978-3-540-32691-5. DOI: 10.1007/978-3-540-32691-5_16. URL: https://doi.org/10.1007/978-3-540-32691-5_16.

[15] D. M. Blei and J. D. Lafferty. "Dynamic Topic Models." In: *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06. Pittsburgh, Pennsylvania, USA: ACM, 2006, pp. 113–120. ISBN: 1-59593-383-2. DOI: 10.1145/1143844.1143859. URL: http://doi.acm.org/10.1145/1143844.1143859.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan. "Latent Dirichlet Allocation." In: *J. Mach. Learn. Res.* 3 (Mar. 2003), pp. 993–1022. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=944919.944937.

[17] S. Büttcher, C. Clarke, and G. V. Cormack. *Information Retrieval: Implementing and Evaluating Search Engines*. The MIT Press, 2010. ISBN: 0262026511, 9780262026512.

[18] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord. "Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings." In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. SIGIR '18. Ann Arbor, MI, USA: ACM, 2018, pp. 35–44. ISBN: 978-1-4503-5657-2. DOI: 10.1145/3209978.3210036. URL: http://doi.acm.org/10.1145/3209978.3210036.

[19] I. Chami, Y. Tamaazousti, and H. Le Borgne. "AMECON: Abstract Meta-Concept Features for Text-Illustration." In: *International Conference on Multimedia Retrieval*. ICMR. 2017.

[20] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. "Large Scale Online Learning of Image Similarity Through Ranking." In: *J. Mach. Learn. Res.* 11 (Mar. 2010), pp. 1109–1135. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1756006.1756042.

[21] K. Chen, T. Bui, C. Fang, Z. Wang, and R. Nevatia. "AMC: Attention Guided Multi-modal Correlation Learning for Image Search." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 6203–6211. ISBN: 978-1-5386-0457-1. DOI: 10.1109/CVPR.2017.657. URL: https://doi.org/10.1109/CVPR.2017.657.

[22] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. "NUS-WIDE: A Real-World Web Image Database from National University of Singapore." In: *Proc. of ACM Conf. on Image and Video Retrieval (CIVR'09)*. Santorini, Greece., July 8-10, 2009.

[23] J. Duchi, E. Hazan, and Y. Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization." In: *J. Mach. Learn. Res.* 12 (July 2011), pp. 2121–2159. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1953048.2021068.

[24] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. "VSE++: Improving Visual-Semantic Embeddings with Hard Negatives." In: *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*. BMVA Press, 2018, p. 12. URL: http://bmvc2018.org/contents/papers/0344.pdf.

[25] M. Fan, W. Wang, P. Dong, L. Han, R. Wang, and G. Li. "Cross-media Retrieval by Learning Rich Semantic Embeddings of Multimedia." In: *Proceedings of the 2017 ACM on Multimedia Conference*. MM '17. Mountain View, California, USA: ACM, 2017, pp. 1698–1706. ISBN: 978-1-4503-4906-2. DOI: 10.1145/3123266.3123369. URL: http://doi.acm.org/10.1145/3123266.3123369.

[26] M. Fedoryszak, B. Frederick, V. Rajaram, and C. Zhong. "Real-Time Event Detection on Social Data Streams." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 2774–2782. ISBN: 9781450362016. DOI: 10.1145/3292500.3330689. URL: https://doi.org/10.1145/3292500.3330689.

[27] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. "What do we perceive in a glance of a real-world scene?" In: *Journal of vision* 7 1 (2007), p. 10.

[28] F. Feng, X. Wang, and R. Li. "Cross-modal Retrieval with Correspondence Autoencoder." In: *Proceedings of the 22Nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, 2014, pp. 7–16. ISBN: 978-1-4503-3063-3. DOI: 10.1145/2647868.2654902. URL: http://doi.acm.org/10.1145/2647868.2654902.

[29] Y. Feng and M. Lapata. "Topic Models for Image Annotation and Text Illustration." In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 831–839. ISBN: 1-932432-65-5. URL: http://dl.acm.org/citation.cfm?id=1857999.1858124.

[30] M. Ferguson, R. Ak, Y. T. Lee, and K. H. Law. "Automatic localization of casting defects with convolutional neural networks." In: *2017 IEEE International Conference on Big Data (Big Data)*. Dec. 2017, pp. 1726–1735. DOI: 10.1109/BigData.2017.8258115.

[31] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. "DeViSE: A Deep Visual-Semantic Embedding Model." In: *Neural Information Processing Systems (NIPS)*. 2013.

[32] R. Girshick. "Fast R-CNN." In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1440–1448. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.169. URL: http://dx.doi.org/10.1109/ICCV.2015.169.

[33] R. Girshick, J. Donahue, T. Darrell, and J. Malik. "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation." In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 580–587. ISBN: 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.81. URL: https://doi.org/10.1109/CVPR.2014.81.

[34] X. Glorot and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks." In: *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Y. W. Teh and M. Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 249–256. URL: http://proceedings.mlr.press/v9/glorot10a.html.

[35] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. "A Multi-View Embedding Space for Modeling Internet Images, Tags, and Their Semantics." In: *Int. J. Comput. Vision* 106.2 (Jan. 2014), pp. 210–233. ISSN: 0920-5691. DOI: 10.1007/s11263-013-0658-4. URL: http://dx.doi.org/10.1007/s11263-013-0658-4.

[36] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[37] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. The MIT Press, 2016. ISBN: 0262035618, 9780262035613.

[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets." In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680. URL: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

[39] R. Hadsell, S. Chopra, and Y. LeCun. "Dimensionality Reduction by Learning an Invariant Mapping." In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2. June 2006, pp. 1735–1742. DOI: 10.1109/CVPR.2006.100.

[40] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing* 2016 (2016), pp. 2116–2121.

[41] W. L. Hamilton, J. Leskovec, and D. Jurafsky. "Diachronic word embeddings reveal statistical laws of semantic change." In: *arXiv preprint arXiv:1605.09096* (2016).

[42] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. "Canonical Correlation Analysis: An Overview with Application to Learning Methods." In: *Neural Computation* 16.12 (Dec. 2004), pp. 2639–2664. ISSN: 0899-7667. DOI: 10.1162/0899766042321814.

[43] K. He, G. Gkioxari, P. Dollár, and R. Girshick. "Mask R-CNN." In: *2017 IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017, pp. 2980–2988. DOI: 10.1109/ICCV.2017.322.

[44] K. He, X. Zhang, S. Ren, and J. Sun. "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification." In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1026–1034. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.123. URL: http://dx.doi.org/10.1109/ICCV.2015.123.

[45] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.

[46] R. Herbrich, T. Graepel, and K. Obermayer. "Large Margin Rank Boundaries for Ordinal Regression." In: *Advances in Large Margin Classifiers*. Ed. by P. J. Bartlett, B. Schölkopf, D. Schuurmans, and A. J. Smola. MIT Press, 2000, pp. 115–132.

[47] A. Hermans*, L. Beyer*, and B. Leibe. "In Defense of the Triplet Loss for Person Re-Identification." In: *arXiv preprint arXiv:1703.07737* (2017).

[48] E. Hoffer and N. Ailon. "Deep Metric Learning Using Triplet Network." In: *Similarity-Based Pattern Recognition*. Ed. by A. Feragen, M. Pelillo, and M. Loog. Cham: Springer International Publishing, 2015, pp. 84–92. ISBN: 978-3-319-24261-3.

[49] H. Hotelling. "Relations Between Two Sets of Variates." In: *Biometrika* 28.3/4 (Dec. 1936), pp. 321–377. ISSN: 00063444. DOI: 10.2307/2333955. URL: http://dx.doi.org/10.2307/2333955.

[50] D. Hu, X. Li, and X. Lu. "Temporal Multimodal Learning in Audiovisual Speech Recognition." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, pp. 3574–3582. DOI: 10.1109/CVPR.2016.389.

[51] J. Hu, J. Lu, and Y.-P. Tan. "Discriminative Deep Metric Learning for Face Verification in the Wild." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2014.

[52] X. Huang, Y. Peng, and M. Yuan. "MHTN: Modal-adversarial Hybrid Transfer Network for Cross-modal Retrieval." In: *CoRR* abs/1708.04308 (2017). arXiv: 1708.04308. URL: http://arxiv.org/abs/1708.04308.

[53] S. Ioffe and C. Szegedy. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift." In: *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*. ICML'15. Lille, France: JMLR.org, 2015, pp. 448–456. URL: http://dl.acm.org/citation.cfm?id=3045118.3045167.

[54] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan. "Learning Consistent Feature Representation for Cross-Modal Multimedia Retrieval." In: *IEEE Transactions on Multimedia* 17.3 (Mar. 2015), pp. 370–381. ISSN: 1520-9210. DOI: 10.1109/TMM.2015.2390499.

[55] C. Kang, S. Liao, Z. Li, Z. Cao, and G. Xiong. "Learning Deep Semantic Embeddings for Cross-Modal Retrieval." In: *Proceedings of the Ninth Asian Conference on Machine Learning*. Ed. by M.-L. Zhang and Y.-K. Noh. Vol. 77. Proceedings of Machine Learning Research. PMLR, 15–17 Nov 2017, pp. 471–486. URL: http://proceedings.mlr.press/v77/kang17a.html.

[56] A. Karpathy and L. Fei-Fei. "Deep Visual-Semantic Alignments for Generating Image Descriptions." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (Apr. 2017), pp. 664–676. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2598339. URL: https://doi.org/10.1109/TPAMI.2016.2598339.

[57] G. Kim, S. Moon, and L. Sigal. "Joint Photo Stream and Blog Post Summarization and Exploration." In: *28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*. 2015.

[58] G. Kim and E. P. Xing. "Time-sensitive Web Image Ranking and Retrieval via Dynamic Multi-task Regression." In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy: ACM, 2013, pp. 163–172. ISBN: 978-1-4503-1869-3. DOI: 10.1145/2433396.2433417. URL: http://doi.acm.org/10.1145/2433396.2433417.

[59] G. Kim and E. P. Xing. "Reconstructing Storyline Graphs for Image Recommendation from Web Community Photos." In: *27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*. 2014.

[60] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." In: *CoRR* abs/1412.6980 (2014). arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980.

[61] R. Kiros, R. Salakhutdinov, and R. S. Zemel. "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models." In: *ArXiv* abs/1411.2539 (2014).

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[63] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems, NIPS 2012*. 2012, pp. 1097–1105.

[64] B. Kulis et al. "Metric learning: A survey." In: *Foundations and Trends® in Machine Learning* 5.4 (2013), pp. 287–364.

[65] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. "Statistically Significant Detection of Linguistic Change." In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Florence, Italy: International World Wide Web Conferences Steering Committee, 2015, pp. 625–635. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741627. URL: https://doi.org/10.1145/2736277.2741627.

[66] J. H. Lau, N. Collier, and T. Baldwin. "On-line Trend Analysis with Topic Models: #twitter Trends Detection Topic Model Online." In: *Proceedings of COLING 2012*. Mumbai, India: The COLING 2012 Organizing Committee, 2012, pp. 1519–1534. URL: http://aclweb.org/anthology/C12-1093.

[67] Q. Le and T. Mikolov. "Distributed Representations of Sentences and Documents." In: *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*. ICML'14. Beijing, China: JMLR.org, 2014, pp. II-1188–II-1196. URL: http://dl.acm.org/citation.cfm?id=3044805.3045025.

[68] Y. J. Lee, A. A. Efros, and M. Hebert. "Style-Aware Mid-level Representation for Discovering Visual Connections in Space and Time." In: *Proceedings of the 2013 IEEE International Conference on Computer Vision*. ICCV '13. Washington, DC, USA: IEEE Computer Society, 2013, pp. 1857–1864. ISBN: 978-1-4799-2840-8. DOI: 10.1109/ICCV.2013.233. URL: http://dx.doi.org/10.1109/ICCV.2013.233.

[69] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. "Multimedia Content Processing Through Cross-modal Association." In: *Proceedings of the Eleventh ACM International Conference on Multimedia*. MULTIMEDIA '03. Berkeley, CA, USA: ACM, 2003, pp. 604–611. ISBN: 1-58113-722-2. DOI: 10.1145/957013.957143. URL: http://doi.acm.org/10.1145/957013.957143.

[70] R. Li and J. Jia. "Visual Question Answering with Question Representation Update (QRU)." In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 4662–4670. ISBN: 9781510838819.

[71] S. Li, W. Zhang, and A. B. Chan. "Maximum-Margin Structured Learning with Deep Networks for 3D Human Pose Estimation." In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 2848–2856. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.326. URL: https://doi.org/10.1109/ICCV.2015.326.

[72] T.-Y. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. "Feature Pyramid Networks for Object Detection." In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 936–944.

[73] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia. "Path Aggregation Network for Instance Segmentation." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 8759–8768.

[74] W. Liu, A. Rabinovich, and A. C. Berg. "ParseNet: Looking Wider to See Better." In: *CoRR* abs/1506.04579 (2015). arXiv: 1506.04579. URL: http://arxiv.org/abs/1506.04579.

[75] X. Liu and B. Huet. "Heterogeneous features and model selection for event-based media classification." In: *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*. ACM. 2013, pp. 151–158.

[76] J. Lu, J. Yang, D. Batra, and D. Parikh. "Hierarchical Question-image Co-attention for Visual Question Answering." In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. Barcelona, Spain: Curran Associates Inc., 2016, pp. 289–297. ISBN: 978-1-5108-3881-9. URL: http://dl.acm.org/citation.cfm?id=3157096.3157129.

[77] L. van der Maaten and G. Hinton. "Visualizing Data using t-SNE." In: *Journal of Machine Learning Research* 9 (2008), pp. 2579–2605. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[78] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge, UK: Cambridge University Press, 2008. ISBN: 978-0-521-86571-5. URL: http://nlp.stanford.edu/IR-book/information-retrieval-book.html.

[79] *MathWorks*. https://www.mathworks.com/discovery/convolutional-neural-network.html. Accessed: 2018-02-08.

[80] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. "Building a Large-Scale Corpus for Evaluating Event Detection on Twitter." In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*. CIKM '13. San Francisco, California, USA: Association for Computing Machinery, 2013, pp. 409–418. ISBN: 9781450322638. DOI: 10.1145/2505515.2505695. URL: https://doi.org/10.1145/2505515.2505695.

[81] A. J. McMinn, Y. Moshfeghi, and J. M. Jose. "Building a Large-scale Corpus for Evaluating Event Detection on Twitter." In: *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*. CIKM '13. San Francisco, California, USA: ACM, 2013, pp. 409–418. ISBN: 978-1-4503-2263-8. DOI: 10.1145/2505515.2505695. URL: http://doi.acm.org/10.1145/2505515.2505695.

[82] P. J. McParlane and J. M. Jose. "Exploiting Time in Automatic Image Tagging." In: *Proceedings of the 35th European Conference on Advances in Information Retrieval*. ECIR'13. Moscow, Russia: Springer-Verlag, 2013, pp. 520–531. ISBN: 978-3-642-36972-8. DOI: 10.1007/978-3-642-36973-5_44. URL: http://dx.doi.org/10.1007/978-3-642-36973-5_44.

[83] T. Mikolov, K. Chen, G. Corrado, and J. Dean. "Efficient Estimation of Word Representations in Vector Space." In: *CoRR* abs/1301.3781 (2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781.

[84] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. "Distributed Representations of Words and Phrases and Their Compositionality." In: *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. NIPS'13. Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119. URL: http://dl.acm.org/citation.cfm?id=2999792.2999959.

[85] N. C. Mithun, J. Li, F. Metze, and A. K. Roy-Chowdhury. "Learning Joint Embedding with Multimodal Cues for Cross-Modal Video-Text Retrieval." In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ICMR '18. Yokohama, Japan: ACM, 2018, pp. 19–27. ISBN: 978-1-4503-5046-4. DOI: 10.1145/3206025.3206064. URL: http://doi.acm.org/10.1145/3206025.3206064.

[86] K. P. Murphy. *Machine learning : a probabilistic perspective*. Cambridge, Mass. [u.a.]: MIT Press, 2013. ISBN: 9780262018029 0262018020. URL: https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2.

[87] V. Nair and G. E. Hinton. "Rectified Linear Units Improve Restricted Boltzmann Machines." In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML'10. Haifa, Israel: Omnipress, 2010, pp. 807–814. ISBN: 978-1-60558-907-7. URL: %5Curl%7Bhttp://dl.acm.org/citation.cfm?id=3104322.3104425%7D.

[88] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. "Multimodal Deep Learning." In: *Proceedings of the 28th International Conference on International Conference on Machine Learning*. ICML'11. Bellevue, Washington, USA: Omnipress, 2011, pp. 689–696. ISBN: 978-1-4503-0619-5. URL: http://dl.acm.org/citation.cfm?id=3104482.3104569.

[89] H. Noh, S. Hong, and B. Han. "Learning Deconvolution Network for Semantic Segmentation." In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1520–1528. ISBN: 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.178. URL: http://dx.doi.org/10.1109/ICCV.2015.178.

[90] Y. Pan, Y. Li, T. Yao, T. Mei, H. Li, and Y. Rui. "Learning Deep Intrinsic Video Representation by Exploring Temporal Coherence and Graph Structure." In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA: AAAI Press, 2016, pp. 3832–3838. ISBN: 978-1-57735-770-4. URL: http://dl.acm.org/citation.cfm?id=3061053.3061155.

[91] Y. Peng, X. Huang, and Y. Zhao. "An Overview of Cross-media Retrieval: Concepts, Methodologies, Benchmarks and Challenges." In: *IEEE Transactions on Circuits and Systems for Video Technology* PP.99 (2017), pp. 1–1. ISSN: 1051-8215. DOI: 10.1109/TCSVT.2017.2705068.

[92] Y. Peng, J. Qi, X. Huang, and Y. Yuan. "CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network." In: *IEEE Transactions on Multimedia* 20.2 (Feb. 2018), pp. 405–420. ISSN: 1520-9210. DOI: 10.1109/TMM.2017.2742704.

[93] Y. Peng, X. Huang, and J. Qi. "Cross-media Shared Representation by Hierarchical Learning with Multiple Deep Networks." In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. IJCAI'16. New York, New York, USA: AAAI Press, 2016, pp. 3846–3853. ISBN: 978-1-57735-770-4. URL: http://dl.acm.org/citation.cfm?id=3061053.3061157.

[94] J. Pennington, R. Socher, and C. Manning. "Glove: Global Vectors for Word Representation." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. URL: http://aclweb.org/anthology/D14-1162.

[95] M. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. "Deep Contextualized Word Representations." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. URL: https://www.aclweb.org/anthology/N18-1202.

[96] M. C. Potter, B. Wyble, C. E. Hagmann, and E. McCourt. "Detecting meaning in RSVP at 13 ms per picture." In: 76 (2013), pp. 270–279.

[97] V. Ranjan, N. Rasiwasia, and C. V. Jawahar. "Multi-label Cross-Modal Retrieval." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 4094–4102. DOI: 10.1109/ICCV.2015.466.

[98] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. "Collecting Image Annotations Using Amazon's Mechanical Turk." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. CSLDAMT '10. Los Angeles, California: Association for Computational Linguistics, 2010, pp. 139–147. URL: http://dl.acm.org/citation.cfm?id=1866696.1866717.

[99] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. R. Lanckriet, R. Levy, and N. Vasconcelos. "A New Approach to Cross-modal Multimedia Retrieval." In: *Proceedings of the 18th ACM International Conference on Multimedia*. MM '10. Firenze, Italy: ACM, 2010, pp. 251–260. ISBN: 978-1-60558-933-6. DOI: 10.1145/1873951.1873987. URL: http://doi.acm.org/10.1145/1873951.1873987.

[100] S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 91–99. URL: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf.

[101] A. Rosenfeld and K. Erk. "Deep Neural Models of Semantic Shift." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 474–484. DOI: 10.18653/v1/N18-1044. URL: http://aclweb.org/anthology/N18-1044.

[102] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge." In: *Int. J. Comput. Vision* 115.3 (Dec. 2015), pp. 211–252. ISSN: 0920-5691. DOI: 10.1007/s11263-015-0816-y. URL: http://dx.doi.org/10.1007/s11263-015-0816-y.

[103] T. Sakaki, M. Okazaki, and Y. Matsuo. "Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors." In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. Raleigh, North Carolina, USA: ACM, 2010, pp. 851–860. ISBN: 978-1-60558-799-8. DOI: 10.1145/1772690.1772777. URL: http://doi.acm.org/10.1145/1772690.1772777.

[104] A. Santoro, D. Raposo, D. G. Barrett, M. Malinowski, R. Pascanu, P. Battaglia, and T. Lillicrap. "A simple neural network module for relational reasoning." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017, pp. 4967–4976. URL: http://papers.nips.cc/paper/7082-a-simple-neural-network-module-for-relational-reasoning.pdf.

[105] P. H. Schönemann. "A generalized solution of the orthogonal procrustes problem." In: *Psychometrika* 31.1 (Mar. 1966), pp. 1–10. ISSN: 1860-0980. DOI: 10.1007/BF02289451. URL: https://doi.org/10.1007/BF02289451.

[106] F. Schroff, D. Kalenichenko, and J. Philbin. "FaceNet: A unified embedding for face recognition and clustering." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682.

[107] D. Semedo and J. Magalhaes. "Temporal Cross-Media Retrieval with Soft-Smoothing." In: *2018 ACM Multimedia Conference on Multimedia Conference*. MM '18. Seoul, Republic of Korea: ACM, 2018, pp. 1038–1046. ISBN: 978-1-4503-5665-7. DOI: 10.1145/3240508.3240665. URL: http://doi.acm.org/10.1145/3240508.3240665.

[108] E. Shelhamer, J. Long, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation." In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4 (Apr. 2017), pp. 640–651. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2016.2572683. URL: https://doi.org/10.1109/TPAMI.2016.2572683.

[109] K. Simonyan and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In: *CoRR* abs/1409.1556 (2014). arXiv: 1409.1556. URL: http://arxiv.org/abs/1409.1556.

[110] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. "Deep Metric Learning via Lifted Structured Feature Embedding." In: *Computer Vision and Pattern Recognition (CVPR)*. 2016.

[111] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *Journal of Machine Learning Research* 15 (2014), pp. 1929–1958. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[112] N. Srivastava, E. Mansimov, and R. Salakhudinov. "Unsupervised Learning of Video Representations using LSTMs." In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 843–852. URL: http://proceedings.mlr.press/v37/srivastava15.html.

[113] N. Srivastava and R. Salakhutdinov. "Learning representations for multimodal data with deep belief nets." In: *In International Conference on Machine Learning Workshop*. 2012.

[114] S. Sukhbaatar, a. szlam arthur, J. Weston, and R. Fergus. "End-To-End Memory Networks." In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett. Curran Associates, Inc., 2015, pp. 2440–2448. URL: http://papers.nips.cc/paper/5846-end-to-end-memory-networks.pdf.

[115] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. "Going Deeper with Convolutions." In: *Computer Vision and Pattern Recognition (CVPR)*. 2015. URL: http://arxiv.org/abs/1409.4842.

[116] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. "DeepFace: Closing the Gap to Human-Level Performance in Face Verification." In: *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 1701–1708. ISBN: 978-1-4799-5118-5.

DOI: 10.1109/CVPR.2014.220. URL: http://dx.doi.org/10.1109/CVPR.2014.220.

[117] M. R. Trad, A. Joly, and N. Boujemaa. "Large Scale Visual-Based Event Matching." In: *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ICMR '11. Trento, Italy: Association for Computing Machinery, 2011. ISBN: 9781450303361. DOI: 10.1145/1991996.1992049. URL: https://doi.org/10.1145/1991996.1992049.

[118] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. "Large Margin Methods for Structured and Interdependent Output Variables." In: *J. Mach. Learn. Res.* 6 (Dec. 2005), pp. 1453–1484. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=1046920.1088722.

[119] M. Tsytsarau, T. Palpanas, and M. Castellanos. "Dynamics of News Events and Social Media Reaction." In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '14. New York, New York, USA: ACM, 2014, pp. 901–910. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623670. URL: http://doi.acm.org/10.1145/2623330.2623670.

[120] T. Uricchio, L. Ballan, M. Bertini, and A. Del Bimbo. "Evaluating Temporal Information for Social Image Annotation and Retrieval." In: *Image Analysis and Processing – ICIAP 2013*. Ed. by A. Petrosino. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 722–732. ISBN: 978-3-642-41181-6.

[121] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen. "Adversarial Cross-Modal Retrieval." In: *Proceedings of the 2017 ACM on Multimedia Conference*. MM '17. Mountain View, California, USA: ACM, 2017, pp. 154–162. ISBN: 978-1-4503-4906-2. DOI: 10.1145/3123266.3123326. URL: http://doi.acm.org/10.1145/3123266.3123326.

[122] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. "Joint Feature Selection and Subspace Learning for Cross-Modal Retrieval." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.10 (Oct. 2016), pp. 2010–2023. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2015.2505311.

[123] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang. "A Comprehensive Survey on Cross-modal Retrieval." In: *CoRR* abs/1607.06215 (2016).

[124] L. Wang, Y. Li, and S. Lazebnik. "Learning Deep Structure-Preserving Image-Text Embeddings." In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016, pp. 5005–5013. DOI: 10.1109/CVPR.2016.541.

[125] X. Wang, Y. Hua, E. Kodirov, G. Hu, R. Garnier, and N. M. Robertson. "Ranked List Loss for Deep Metric Learning." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[126] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan. "Cross-Modal Retrieval With CNN Visual Features: A New Baseline." In: *IEEE Transactions on Cybernetics* (2016).

[127] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. "A Discriminative Feature Learning Approach for Deep Face Recognition." In: *Computer Vision – ECCV 2016*. Ed. by B. Leibe, J. Matas, N. Sebe, and M. Welling. Cham: Springer International Publishing, 2016, pp. 499–515. ISBN: 978-3-319-46478-7.

[128] D. R. Wilson and T. R. Martinez. "The General Inefficiency of Batch Training for Gradient Descent Learning." In: *Neural Networks* 16.10 (Dec. 2003), pp. 1429–1451. ISSN: 0893-6080. DOI: 10.1016/S0893-6080(03)00138-2. URL: https://doi.org/10.1016/S0893-6080(03)00138-2.

[129] C.-Y. Wu, R. Manmatha, A. J. Smola, and P. Krahenbuhl. "Sampling Matters in Deep Embedding Learning." In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2017.

[130] Y. Wu, S. Wang, W. Zhang, and Q. Huang. "Online low-rank similarity function learning with adaptive relative margin for cross-modal retrieval." In: *2017 IEEE International Conference on Multimedia and Expo (ICME)*. July 2017, pp. 823–828. DOI: 10.1109/ICME.2017.8019528.

[131] Y. Wu, S. Wang, and Q. Huang. "Learning Semantic Structure-preserved Embeddings for Cross-modal Retrieval." In: *Proceedings of the 26th ACM International Conference on Multimedia*. MM '18. Seoul, Republic of Korea: ACM, 2018, pp. 825–833. ISBN: 978-1-4503-5665-7. DOI: 10.1145/3240508.3240521. URL: http://doi.acm.org/10.1145/3240508.3240521.

[132] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, and Z. Huang. "Modal-adversarial Semantic Learning Network for Extendable Cross-modal Retrieval." In: *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. ICMR '18. Yokohama, Japan: ACM, 2018, pp. 46–54. ISBN: 978-1-4503-5046-4. DOI: 10.1145/3206025.3206033. URL: http://doi.acm.org/10.1145/3206025.3206033.

[133] F. Yan and K. Mikolajczyk. "Deep correlation for matching images and text." In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2015, pp. 3441–3450. DOI: 10.1109/CVPR.2015.7298966.

[134] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, and J. Luo. "Deep Multimodal Representation Learning from Temporal Data." In: *CVPR*. IEEE Computer Society, 2017, pp. 5066–5074.

[135] T. Yao, T. Mei, and C. W. Ngo. "Learning Query and Image Similarities with Ranking Canonical Correlation Analysis." In: *2015 IEEE International Conference on Computer Vision (ICCV)*. Dec. 2015, pp. 28–36. DOI: 10.1109/ICCV.2015.12.

[136] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. "Dynamic Word Embeddings for Evolving Semantic Discovery." In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. WSDM '18. Marina Del Rey, CA, USA: ACM, 2018, pp. 673–681. ISBN: 978-1-4503-5581-0. DOI: 10.1145/3159652.3159703. URL: http://doi.acm.org/10.1145/3159652.3159703.

[137] Z. Yao, Y. Sun, W. Ding, N. Rao, and H. Xiong. "Dynamic Word Embeddings for Evolving Semantic Discovery." In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, Marina Del Rey, CA, USA, February 5-9, 2018*. Ed. by Y. Chang, C. Zhai, Y. Liu, and Y. Maarek. ACM, 2018, pp. 673–681. DOI: 10.1145/3159652.3159703. URL: http://doi.acm.org/10.1145/3159652.3159703.

[138] D. Yi, Z. Lei, S. Liao, and S. Z. Li. "Deep Metric Learning for Person Re-identification." In: *2014 22nd International Conference on Pattern Recognition*. Aug. 2014, pp. 34–39. DOI: 10.1109/ICPR.2014.16.

[139] M. D. Zeiler. "ADADELTA: An Adaptive Learning Rate Method." In: *CoRR* abs/1212.5701 (2012). arXiv: 1212.5701. URL: http://arxiv.org/abs/1212.5701.

[140] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. "From Recognition to Cognition: Visual Commonsense Reasoning." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2019.

[141] X. Zhai, Y. Peng, and J. Xiao. "Learning Cross-Media Joint Representation With Sparse and Semi-supervised Regularization." In: *IEEE Transactions on Circuits and Systems for Video Technology* 24.6 (June 2014), pp. 965–978. ISSN: 1051-8215. DOI: 10.1109/TCSVT.2013.2276704.

[142] H. Zhang and L. Chen. "Learning optimal data representation for cross-media retrieval." In: *2012 19th IEEE International Conference on Image Processing*. Sept. 2012, pp. 1925–1928. DOI: 10.1109/ICIP.2012.6467262.

[143] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal. "Context Encoding for Semantic Segmentation." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. June 2018, pp. 7151–7160. DOI: 10.1109/CVPR.2018.00747.

[144] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. "Single-Shot Refinement Neural Network for Object Detection." In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.

[145] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling. "M2Det: A Single-Shot Object Detector Based on Multi-Level Feature Pyramid Network." In: *AAAI*. 2018.